NOTE TO USERS

THE METACAUSAL THEORY OF AUTONOMY

by

RICCARDO C. REPETTI

A dissertation submitted to the Graduate Faculty in philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2005

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the
Graduate Faculty in Philosophy in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

May 4, 2005
Date

Chair of Examining Committee

May 4, 2005
Date

Executive Officer

John Greenwood

Stephen Grover

Michael Levin
Supervision Committee

Harry Frankfurt

Galen Strawson
Examiners

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE METACAUSAL THEORY OF AUTONOMY

by

RICCARDO C. REPETTI

Advisor: Professor Michael Levin

David Chalmers (1996) distinguishes hard/easy problems of consciousness, identifying hard ones as the puzzling metaphysical issues, and easy problems as involving specific causal/functional relations between consciousness and brain/behavioral states. This shifts metaphysical concerns to the speculative background, and brings neurophilosophical ones to the fore. I distinguish hard/easy problems of autonomy, identifying hard ones as puzzling metaphysical issues, and easy problems as involving specific causal/functional relations between autonomy and brain/behavioral states. To solve the easy autonomy problem, I apply a causal/functional analysis to Frankfurt's (1971) meta-motivational model, yielding a model of autonomy as metacognitive self-regulative or "metacausal" control. Since intrinsically-causal, the account is intrinsically compatible with determinism. It also handles all the nuances in Frankfurt-cases involving counterfactual interveners, grounds a determinism-friendly version of "PAP", the principle of alternate possibilities, and provides an error theory for the inflated intuitions of contracausalists. The metacausal theory has the apparatus to establish easy-style meanings for such disputed notions as 'possibility' and 'ability', breaking traditional stalemates. The "consequence argument" claims determinism entails we have no free will, but I argue that it contains an implicit commitment to "deterministic actualism" that is self-refuting, it commits two modal fallacies, serious fixed/variable and attributive/referential equivocations, and other serious flaws.

Table of Contents

Chapter One: General Introduction

Analysis of our daily experience – its concepts, language, and phenomenology – as agents deliberating, choosing, acting, and appraising supports the ubiquity of freedom, and folk platitudes embed the notion deep within our conceptual-linguistic scheme. Since context governs intuition, we cannot divorce freedom from ordinary experience, nor ignore the phenomenology of choice. All our normative judgements, reactive attitudes, and normative attributive practices presuppose that we control and are responsible for our actions (Strawson 1962). We excuse people to the extent that they lack self-control, e.g., if mentally ill. Outside philosophy, talk of freedom arises only in its absence.

Within philosophy, *pessimism* about freedom is due mostly to "determinism",[1] the doctrine that every event is the lawful result of a cause, and comes in three forms: "Close-range" doubts we control actions (freedom of action); "middle-distance" doubts we control our wills (freedom of will); and "pessimism at the horizon" doubts we have ultimate control over ourselves, since we lack control over self-formation (freedom of self) (Russell, 2002). Determinism is usually seen as the culprit, though variants are God and logic: If God knows all truths, or if true propositions are timelessly true, the future does not seem open.

I am *optimistic* about all three forms of freedom. Mele (1995), an agnostic optimist, claims autonomy is more credible than nonautonomy, whether determinism is true or not, as do I; but I promote soft determinism. I argue that metaphysical freedom is compatible with determinism ("compatibilism"), and is found in the hierarchical will's (Frankfurt, 1971) causal control or "metacausal autonomy". My use of the "meta-" prefix is standard; e.g., a

---

[1] For Strawson (1962), "pessimists" hold determinism and freedom incompatible, and "optimists" hold them compatible. Russell discusses divergent usage: "the ... pessimist may well be an 'optimist' about ... (libertarian) [freedom]" (2002, n.1, p.253). I call "pessimists" non-autonomists and "optimists" autonomists.

metamotive is a motive about one's motive (Dennett, 1984). I contrast "metaphysical" with normative (Wolf, 1990) and pragmatic (Strawson, 1962) conceptions of freedom, and use "metacausal" to distinguish the sort of higher-order mental causation involved in metacognition from cognitive, non-cognitive, non-mental or purely physical causation. With Searle (1992), I see the mental as higher-order physical, so my ideas are monistic.

Chalmers differentiates *hard* and *easy* problems of consciousness (1996), and identifies the hard ones as accounting for the metaphysics of mind, e.g., its irreducible subjectivity, and the easy problem as just that of giving a causal/functional account of its input/output – "I/O" – operations. Chalmers says we may never resolve the hard problems, but answers to the easy problem are available. This shifts philosophy of mind to the speculative background, and neurophilosophy to the foreground. A like division of issues for freedom views as hard problems metaphysical questions such as whether determinism is true, and views as the easy problem the task of giving a causal/functional account of our apparently autonomous behavior or I/O operations. This shifts the stalemates to the speculative background, and neurophilosophy to the foreground. In this dissertation, I attempt a solution only to the easy problem. This easy-problem analysis of Frankfurt-style, hierarchical-will-type autonomy has the virtue of being intrinsically causal, thus intrinsically deterministic, hence transparently compatibilist.

'Hard' problems involve horizon-level pessimism, and rest on "incompatibilist" presuppositions enshrined in the "consequence argument" ("CON") advocated by van Inwagen (1975), G. Strawson, Pereboom, Honderich, Fischer, Kane, and others.[2]

> If determinism is true, then our acts are the *consequences* of the laws of nature and events in the remote past. But it is not up to us what went on before we were born,

---

[2] All in Kane (2002a), my most cited work, though cites are to the authors directly.

> and neither is it up to us what the laws of nature are. Therefore, the *consequences* of these things (including our present acts) are not up to us. (Van Inwagen, 1983, p. 16; emphases added)

The easy-autonomy approach is opposed by advocates of CON, many of whom typically grant mid- and close-range optimism (e.g., G. Strawson, Pereboom, Honderich, *supra*), but dig their heels in at the horizon with CON-concerns about *ultimate* and thus *real* control. Non-natural optimists – libertarians – agree on CON's defeating ultimate control, so they reject determinism. Thus, a successful attack on CON would displace all such opponents of naturalistic optimism in one move, as I aim to achieve in chapter 2 with my critique of CON.

I attempt to break the *conceptual stalemate* between incompatibilists and compatibilists over "possible", "can", and "free" with a critique of CON that renders "easy-autonomy" plausible. To begin to *explain* easy-autonomy, my causal/functional analysis of metamentality proceeds via a "primatological just-so" story, an *illustrative* analysis of how these abilities *may have* evolved in primates. Simulation may involve an ability to run cognitive/conative mechanisms "off-line", which ability may figure metacausally in the neurophilosophy of autonomy. The causal/functional features of metacausality underpin my theories of the will and freedom, and provide my easy-problem solution.

My view is that agents are autonomous *iff* they have metacausal self control. My model improves on Fischer and Ravizza's ("F&R's") *semi-compatibilism* whereby agents are responsible for their actions *iff* they exhibit "reason-responsive guidance-control" over their "actual-sequence" behavior, though, given determinism, they lack alternate possibilities, so they do not exhibit autonomous "regulative-control" or ability to do otherwise (1998). My model supports determinism-friendly "counterfactual-sequence" regulative-control.

I analyze the problem in folk, phenomenological, and philosophical terms, to trace their progression in the debate. This reveals an *ubiquity* and *heterogeneity* of voluntary

behavior, and so of the data. The data suggests that single-criterion paradigms misrepresent. The debate narrows to a stalemate of differences between two views of conditional ability, of "could have done otherwise": (a) under *identical* conditions ("contracausal" or "acausal"), versus (b) under the *different condition* that one wanted to ("counterfactual"). Each avers the conditional ability to choose/do otherwise *if one wishes*, so is conditional, though "conditional" is associated with (b). But contracausalists allege the ability can operate *contrary to* the causal input, acausalists allege the ability *without any* causal input, and counterfactualists allege the ability *only if one wanted otherwise*, i.e., fully conditionally.

In chapter 2, I allege three unnoticed fallacies in CON. The first involves fixed/variable and attributive/referential equivocations in different stages of the argument; the second, an unjustified homogenizing of necessities in CON's modal weakening rule; and the third, the alleged transfer of prenatal non-control. Analysis of CON reveal a fallacious doctrine that only what is actual or entailed by it is possible. I call this *view of possibilia* "actualism", and argue that *deterministic actualism* is oxymoronic.

These fallacies threaten compatibilist conceptions of *possible, can,* and *control.* Exposing them *breaks the stalemate.* Since "prenatally determined" does not rule out "agent-controlled", and counterfactuals need not be actualized to be licit, compatibilism is the preferred default view, since naturalism is favored over non-naturalism by Ockham's razor, related meta-principles, and the easy-problem approach. So, non-actualistic counterfactuals and the weak, practical possibilities they ground are enough to allay CON's "ultimacy" worries. I argue next that Lehrer's argument (1966) against conditional analyses of *can* equivocates on the capacity/ability distinction. Lastly, I apply the criticisms of CON to fatalistic variants involving God's foreknowledge and timeless truths.

Chapter 3 is modeled on Dennett's (1984) "just-so" evolutionary stories of

consciousness and freedom, according to which they are shown plausible by a sketch of how they *could have* arisen naturally. My just-so story for how metamental autonomy *could have* arisen refers to studies about gaze-following (Baron-Cohen, 1995, pp. 38-43; Bogdan, 2003, pp. 16-17; Butterworth, 1991, 1995; Byrne and Whiten, 1991; Povinelli, 1996; Povinelli and Eddy, 1996; Reynolds, 1993; Tomasello and Call, 1997), mimicry (Bogdan, 1997, 2003; Blackmore, 1999), and other primitive forms of mind-reading (Bogdan, 1997, 2003; Gomez, 1990a, 1990b, 1991, 1994, 1996a, 1996b) that *could have* evolved into simulation, metamentality, and metacausality. A sketch of "simulation theory" is set against one of the "theory-theory" of folk psychology, and a "hybrid theory" is suggested for our lesser purposes. A causal/functional analysis of the simulation mechanism models an ability to run cognitive/conative mechanism "off-line" and that involve metacausation. Analysis of other off-line and meta-phenomena such as language and dreams suggest that an increase in metacausation makes an increase in autonomy *possible*.

Chapter 3 presents an evolutionary just-so story that illustrates the causal/functional properties that make close- and mid-range forms of agential control *possible*. How these elements play these close- and mid-range control roles is spelled out in chapter 4.

Chapter 4 argues that our model provides a solution to the easy-autonomy problem, and grounds a coherent theory of the will itself. Analysis of Frankfurt's (1971) and Ainslie's (2001) theories of the will supplies the basis for adequacy conditions for any theory of the will, and supports the claim that the metacausal theory better satisfies them. Our causal analysis of the will thus supports our answers to mid- and close-range forms of pessimism.

In chapter 5, the theory of autonomy is developed against the competing theories of F&R (1998), Frankfurt (1971), Wolf (1990) and Dennett (2003), and soft compatibilists like Mele (1995). "Frankfurt-cases" use a counterfactual "Intervener" to defeat "PAP", *the*

principle of alternate possibilities, according to which "Agent" is responsible *iff* in acting he has alternate possibilities open to him. Frankfurt-cases stipulate that Agent does what he wants to do, and that Intervener does not intervene, but stands by ready to intervene in case Agent attempts otherwise. Since Intervener removes alternate possibilities, but doesn't causally interfere with Agent's choice, Agent remains responsible. Many accept the Frankfurt-case rejection of PAP, and Fischer (2002) contends that it is so intuitive that a demand for an explication is equivalent to a demand for a definition of jazz. Many adopt PAP-free compatibilism. F&R (1998) develop semi-compatibilism in response: It is semi-compatible because CON removes alternate possibilities and thus autonomy, but leaves room for Frankfurt-style responsibility, which they explicate as actual-sequence guidance-control, however jazzy and undefined, which they contrast with CON-violating regulative-control.

I adduce a determinism-friendly "PAPW" that has two parts. "PAPW1" – jazz defined – explains why Frankfurt-case intuitions support the divorce of responsibility from full autonomy: Agent acts voluntarily *iff* he *would* have done what he did in the nearest possible world in which Agent *had* alternate possibilities. In light of my objections to CON, PAPW1's counterfactual terms are facially consistent with determinism. F&R's analysis of the actual versus the counterfactual sequence of action determines whether Agent exhibits reason-sensitivity and thus guidance-control, but on our model the same analysis is used to reveal determinism-friendly or weak regulative-control, captured in PAPW2. F&R-style cases are used to illustrate that PAPW better handles the issues. It also handles Wolf's *asymmetry problem*, according to which determinism supports praise but not blame (1988, 1990). PAPM grounds desert, thereby eliminating universal exculpation.

Chapter 5 finalizes my responses to mid-/close-range pessimists by specifying the horizon-resistant principles of mid-/close-range control, PAPW and PAPM. These principles

identify genuinely "self-forming acts" or "SFAs" (Kane, 2002b) that are determined *and* ultimacy-generating. This self-formation model escapes the objection that since self-formation is determined, agents lack responsibility for character and thus action. A dilemma is that self-formation requires either an infinite series of prior selves or self-creation *ex nihilo* (Wolf, 1990). These alleged impossibilities are shown possible on my model.

In chapter 6, I tie up loose ends. I also set forth an error theory that explains the inflated intuitions of incompatibilist optimists by reference to the phenomenology of the causal/functional features of the cognitive architecture of autonomy. And I suggest some applications of the metacausal theory.

Believers in freedom are "libertarians" or "soft determinist" compatibilists, but a single term is needed for both. I use "autonomist"; so does Mele (1995).

Philosophical doubt does not trump the presumption that lies with ordinary folks. The burden is on pessimists when intuitions clash about *can* and related notions; revisionary metaphysics is unwarranted. We associate many daily factors with autonomy, so we must attend to *data* before *theory*. "Data" includes simple facts like weighing pros/cons. Some choices need no deliberation; others need lots. Some are made under duress; some require intense struggle. A range of behaviors comprises the *data*. So, *theory-level* attempts to identify *one* ability as paradigmatic are ill-fated: Optimists propose one, and pessimists conjure counterexamples, but this is a red herring. One proposal is *doing what one wants*: Shelley's thirsty, so she drinks water. This is "classical compatibilism", held by Hobbes, Locke, Hume, Moore, Schlick, Ayer, Stevenson, Davidson, etc. But, Frankfurt notes (1971), animals, children, and the mentally ill – paradigms of unfreedom (stipulatively, "non-persons") – enjoy this 'freedom'. Shelley successfully *refrains*, but Barbara binges despite trying to refrain. Since nonpersons more resemble Barbara, *refraining* better captures our

intuitions – a better *theory-data* 'fit'. But *animals also restrain* hunger for safety: A rat approaches food, but retreats from a nearby cat.

Many think the key is *ability to do otherwise*. We refrain *without* countervailing motives; animals only act on the strongest *occurrent* desire. Greater *cognitive* ability enables us to rank desires, but we can also refrain for moral reasons. The moral construal of freedom works in cases where one must choose to do the right thing, despite consequences, but a metaphysical construal makes sense with *amoral* activities like athletics merit praise, which presupposes causal authorship, a key component in the concept of autonomy.

We also act *at will* for its own sake. This is *pure* restraint, without a *particular* reason, for its own sake, an exercise of itself. When questioned about freedom, many often raise or don't raise their hand, to display their alternating power to perform or not perform bodily actions. This shows that freedom is ordinarily construed as an ubiquitous motor ability. It is unconvincing because it only illustrates the ability to raise one's hand *when one wants to* or to not raise it *when one wants not to*, but nonpersons can do these.

That one can *fully* want to raise one's hand at T *and* refrain from raising it *at T* is dubious; raising it when one doesn't want to *at all* is odd. Try the former; it's as hard as to rotate one's right hand clockwise while rotating one's right foot counter-clockwise – not impossible. Those inclined to this Moorean proof construe acting *without motive* as acting *at will*, but 'unintentional action' is oxymoronic if 'action' is intentional; 'behavior' need not be. I *cannot* 'raise' my arm *without wanting to*, though I can *to demonstrate control* over my movements; *the aim generates the ability*. These motor-control 'experiments' suggest that what we are *actually* able to do is not *to refrain* but *to alter* our desires for abstract reasons, only *apparently contra* the Humean postulate against the independent power of reason.

So, when we get the feel going in support of raising our arm, but prevent ourselves,

we shift between opaque pro/con reasons/motives. Pure restraint is ability to engage counter-reasons for their own sake, at will. These abilities provide means to strengthen the will, and there are as many as there are categories of action. Many practice restraint *just to strengthen the power of restraint.* A person might intentionally weaken his will, alternately.

Since the will's aim is *the will's strengthening/weakening*, this is "*meta*teleological". Pure restraint is but one metateleological motive or "metamotive". Metamotives, Frankfurt (1971) argues, are keys to autonomy. It is no problem that metamotives are caused by conditions stretching into the distant past. Rather, it is problematic to claim that causation is suspended here, as if, when we performed act A under meta-circumstances C, we still could have done otherwise *under C.* Both contracausal conceptions entail that though Shelley drank the water under C, she could have chosen not to under C.

What makes the effective desire effective? It cannot be merely the tautology that "the winning desire is the one that wins". Perhaps "strongest desire" means the desire with the greatest proprioceptive phenomenology. But the effective desire is often *not* the one with the greatest phenomenological force, but rather one we choose because it is in our best interests overall, meta-interests included. The "unconstrained desire" view is the Humean view that freedom is acting on unconstrained desire. The "effective desire" view is the claim that freedom is acting on the phenomenologically strongest desire. The "considered judgment" view is the claim that freedom just is acting on that desire that is most approved by one's considered judgments. Finally, the "metateleological", "metamotivational", or "meta-will" view is the claim that freedom just is acting on "metadesires". Restraint may be a rudimentary form of sophisticated metacausal ability, but only persons act on metadesires – key components of personhood centrally linked with autonomy.

Nonpersons act on strongest desires, but lack judgment/meta-ability. A dog wants to

urinate, but not to be punished, so he doesn't; a deranged man wants to shove someone, but doesn't want to get caught, so he doesn't. Though 'restraint' by stronger desires is present in nonpersons, what distinguishes us is our degree/range of restraint. The punishments are construable as stimuli that elicit the behaviorists' "conditioned response" of self-control, evidencing the utilitarian link between control, response-ability, and responsibility.

Pure restraint acts against strongest motive or for itself; only we do this. It is what a *strong* will is composed of; its privation is akrasia. The athlete strengthens will against internal/external obstacles. Because much goes into it, we attribute successes to his efforts, not *general* causal processes. His achievements did not just *happen*, though being born with predispositions *did*. The battle of self-mastery is an intense doing, no mere happening.

Since kratic/akratic beings (Aristotle, 1925) are caused to be so, hard determinists claim both lack ultimate control. But this ignores the different powers of the continent/incontinent: Only the continent control their bladders. Thus, it is a mistake to blame the incontinent for urinating inappropriately, but not the continent. So, autonomy is more a metaphysical than a normative matter, though normative work may build a bridge from the metaphysics to the ethics of action. Though effort to become an athlete is determined, its phenomenology must be taken into account. Most professors grade students not for *capacities*, but for *exercised demonstration* of such in professor-structured performances requiring efforts even in "natural geniuses". These considerations ground normative judgments, for we praise people because they deserve it from their doings – not because events happen to them, but because of self-controlling abilities.

Restraint produces another effective desire. Although tautological, whichever desire we make effective wins out. So, we cannot identify any ineffective desire as one that was 'supposed' to be effective, even if after tremendous struggle, since struggle shows that the

effective desire was a strong motivating force. Sometimes reason is stronger than passion. Intelligence enables us to deliberate and form reasons for restraint that diverge from desire. The mentally ill display intelligence without prudence, and prudential *insight* is no guarantee against akrasia. The will is thus distinct from practical and theoretical reason. Restraint is not coextensive with autonomy.

Many think ability to choose otherwise is needed: The best choice A is not free if Agent cannot do ~A. Say Shelley chose A under C, but could have chosen ~A. But if she could have chosen ~A *because* she wanted ~A, her different want entails ~C. For C must include motivational set M1 that cannot be the same as M2 if a desire to do otherwise is in M2 but not M1. Thus, " under C, Shelley could have chosen ~A" is a radical claim that means she could have done ~A even if she didn't want to do ~A, akin to saying one can raise one's arm without the desire to do so. This is contracausality.

For determinists, C entails A, but on the aforementioned view it does not, so it is incompatibilist; C may occur and ~A may follow. CON is unnecessary to see the incompatibility between this *contracausality* and determinism. Nothing is incompatible about the *counterfactual* claim that Shelley could have done ~A if she had different (M2) reasons for action ("RFAs") ~C. *If ~C, then ~A* is compatible with *if C, then A*, but some think they can do ~A under C. But, since our behavior flows from our reasons, this can mean only the oxymoron, *unintentional action.*

Contracausalists counterclaim that determinism implies predictability, but if presented with clear predictions about arm-raising decisions, Agent can falsify them. But perfect predictions should factor this too. This gets subtle, but persons wanting to falsify predictions will do what they can to falsify them, and agents with such motives have one set M1, and agents lacking these motives have another M2. Left-hand-raising actions require

left-hand-raising desires. Whichever desire one adopts is a function of other factors, like the desire to foil predictions. "Can do otherwise" cannot mean that when one has an effective desire to raise one's left hand, one is able in that precise state to raise one's opposite hand. (Please repeat the motor-control 'experiment'.) That this is freedom is a misimpression.

If one's restraint-including considered judgment is "raise the left arm", but the right one goes up, that would be a neurological malfunction, akin to steering the wheel to the left, only to find the car veering to the right – not exceptional ability, but malfunctioning. Incontinence is analogous, as are some mental illnesses – not 'abilities'. These illustrate that Agent's reasons are causally unrelated to his actions. Contracausal views cannot do justice to this causal relation.

The metacausal theory embraces Frankfurt's (1971) higher-/lower-order motivational states distinction.[3] I propose a *causal* analysis of the differences between cognitive/conative and metacognitive/metaconative states. Cognitive/conative states are basic sensory-receptor/motor-reactor states – exteroceptive or proprioceptive – in which the organism is in stimulus-response ("S/R") contact with its environment. Metacognitive/metaconative states are about other cognitive/conative states, e.g., thinking about thoughts or wanting a desire to be effective. Conative states involve a cognitive feedback-loop; so, too, for "meta-" ones. All motivational states are intentional in Brentano's "aboutness" sense, but the converse is false. If one wishes to separate "meta/cognitive" from "meta/motivational" terms to exclude the latter from the former, "metamental" remains inclusive.

Some identify freedom as a "mesh" between higher-/lower-order motivational-hierarchy tiers. These are "hierarchical" or "mesh" theories. The metacausal theory is

---

[3] Dworkin made the distinction first (1970); Shatz (1986) critically reviews that literature.

hierarchical, but more general in its concern with the larger category of metamental – not metamotivational – states, and their causal features. The metacausal view is that autonomy involves causal/functional control by one's metamental states of one's decisions/actions.

Compare the *meta*causal view of ability with the oxymoronic causal malfunction suggested by contracausal views: The ability to do otherwise under *slightly different* motivational circumstances differs radically from the ability to do otherwise *without motive* or under *identical* circumstances. If one does otherwise *even if one didn't want to*, that equates with the broken-steering-mechanism case. Thus, only the 'lesser' ability to do otherwise *when we want to* is desirable – acting *unintentionally* or *contraintentionally* is not.

Counterfactual ability requires conditions M2 counter to those in place M1. Had M1 differed (M2), one would have done otherwise. *Contracausal* ability permits identical conditions M1. Imagine the cosmic-history on film God can rewind. When rewound to when Shelley chose A under M1, she chooses B, and on the next rerun C, etc. Any pattern other than A's *ad infinitum* invites *contracausality*. The *counterfactual* view requires all cosmic reruns yield A's *ad infinitum* under M1: Shelley always chooses A since she has the same RFAs M1 for A.

Acausality is a species of contracausality, for choosing ~A under C *without motive* is choosing ~A *contra* C anyway. All determinists reject contracausality, though hard determinists and other incompatibilists hold that autonomy requires it. Let me sketch reasons for and against the causal and contracausal views here.

(1) The main reason philosophers are drawn to the causal/counterfactual view is that it connects actions/beliefs/desires in a rational, scientifically-legitimate way: Beliefs/desires *cause* actions in law-like, predictably-rational ways. Since Pierre believes Adele is in Lyons, and wants to see her, he will – *ceteris paribus* – go to Lyons. Beliefs/desires render actions

rational, for different ones have different causal powers that account for different actions, hence identical belief/desire-inputs yield identical action-outcomes.

(2) The main reason philosophers are drawn to contracausality is that it captures the feel of choice. The phenomenology suggests we can raise our hand or not, at will, regardless of belief/desire, a power that elevates us beyond S/R-predictability, robotic cause/effect pairs M1/A in all reruns. If we cannot do otherwise under identical circumstances, we're not responsible for what we do, it doesn't originate from us, and we are mere dominos in an ancient cause/effect series related by laws. But it doesn't feel that way when we choose; experience seems genuinely up to us, so we believe we are ultimate originators of our actions. We cause them, so we are responsible for them.

(1') The counterfactual view suggests choices are remotely determined, so unfree. If identical causes produce identical effects, then identical M-sets produce A-choices, so they are no more free than a billiard ball is 'free' to roll left when struck a certain way.

(2') The contracausal view suggests the ability we deploy in making decisions is not controlled by belief/desire causal forces that feed into them, so it is difficult to imagine how agents can claim authorship of or responsibility for them. If my action is not caused by my M-states, what caused it? If random, what is it about a random force passing through one at choice-moment-T that makes the choice something Agent can claim to originate?

(1') and (2') feed into the pessimistic dilemma:

PD1. Either our choices are caused or they are random.
PD2. If they are caused, then we are robots who lack freedom.
PD3. If they are random, then we cannot claim authorship over our choices and thus we lack freedom.
PD4. Thus, either way, we lack freedom.

PD2 is hard determinism, PD3 is "hard indeterminism", and PD4 is general pessimism. Hard determinists say autonomy requires contracausality – impossible, and so is autonomy;

optimistic contracausalists are "libertarians". This dilemma shapes the current debate.

On another conception, our ability to choose freely – a manifestation of our nonphysical mind – makes us God-like in that we can inject novel causes into the stream of physical events (Chisolm, 1982). Mind-body dualists could welcome the contracausal conception. This suggests one's metaphysics plays a role in one's view of freedom. Some opposing views can be adjudicated only by reference to the merits of competing foundation-level assumptions, say, the closure of physics versus the doctrine of privileged access.

A similar face-off occurs between folk psychology and eliminativism. An argument that takes its lead solely from general considerations might appear strong because its metaphysics is in vogue, and yet be weak because its foundation-level assumptions have little else to recommend them than their roles in the larger belief system. For all we know, consciousness may be pseudo-emergent energy that does not obey *known* physical laws, but may still count as physical, as does electromagnetism. If so, stalemates between physicalism and dualism, balanced between the closure of the *known* physical domain and the phenomenology of consciousness, need to be rethought.

Further, a spectrum of choices ranges from whim to intense struggle. A model of freedom that takes acting on whim as its paradigm will view contracausal ability as paradigmatic; other models may see this as key, e.g., when backward-looking reflection suggests better choices. The whim model makes *phenomenological* sense only where options are equal or unimportant. But where options are unequal or important, whim-power is sub-optimal. Here we want choice to be constrained by rational factors. Luther "could do no other", but that didn't mean he lacked autonomy; rather, it meant he was committed to his stance. It would be irrational for him to exercise brute power to refrain, given the *importance – not strength* – of his desire. If strength *did* propel actions such that we *cannot*

resist them, intuition suggests we *lose* responsibility for them. Thus, what seems intuitive under one paradigm for the data is not under another. Thus, where one is being tracked or manipulated by someone who knows one's motives, survival favors ability to diverge from predicted patterns (Dennett, 1984). Here, ability to do otherwise *despite* one's motivations seems coextensive with freedom.

Let us briefly mention some historical concerns with freedom. Sophocles' Oedipus is prophesied to fall in love with his mother and kill his father; attempts to avert the prophecy secure its outcome. On this form of fatalism, only some events are inevitable. Dennett calls this "local fatalism", and "global fatalism" the view that all events are inevitable (1984). Many nonphilosophers espouse local fatalism; colloquial expressions express it: "It was her time to die". Few philosophers ascribe to local fatalism.

Aristotle rejected "logical fatalism" in his talk of tomorrow's sea battle (1941). Either there will be a certain sea battle tomorrow or there will not be. If there will be, then it is a fact now that, tomorrow, there is a sea battle; if there will not be a sea battle, then it is a fact now that, tomorrow, there is no sea battle. It is a fact that there will be a sea battle or, alternately, that there will be no sea battle; we just don't know yet which one is the fact.

Facts, like truths, when viewed in this way, are timeless and fixed, and there is nothing we can do to change them – before, during, and certainly not after their occurrence. Only time travelers or God challenge the fixity of the past, but on this view the present and future are equally fixed. Nothing can prevent it that what happens tomorrow happens. *Que sera, sera.* Logical fatalists argue that every event is described by a true statement, and all true statements are timelessly true, so they are true 'before' the events mentioned in them occur. If so, they are unalterable, so there is no freedom. But this is a *non-sequitur*, since if a statement is timelessly true, talk about 'when' it is true makes no sense.

Logical fatalism implies that it doesn't matter what we try to do. Suppose a logical fatalist believed this, and had just fallen from a ship and was swept away by powerful currents. He would believe it was already a fact whether he survives, one that has nothing to do with any efforts on his part. But if he believed this, chances are high he would drown. Belief in fatalism thus affects the future, though fatalists would say that belief in fatalism was itself fated, together with loss of motivation. But if belief effects motivation, this contradicts the tenet that beliefs/desires play no role in cosmic history. Fatalism, an often troubling view, is a troubled view as well.

A more complex case involves the tension between freedom and God's omniscience. If God knows all facts timelessly, then God knows Shelley will choose A beforehand. Since God cannot be mistaken, she must choose A, so how can she choose freely? "Theological fatalism" is similar to logical fatalism, but where logical fatalism sees inevitability as a logical/metaphysical matter, theological fatalism sees it as an epistemological function, of God's omniscience. There are similarities between all forms of fatalism and hard determinism, so we will address them together in chapter 2.

Determinism sees every event as determined lawfully – necessitated – by previous ones. So, if we knew the exact location and trajectory of every particle in the universe at any given time – a full description of the universe or "state description" – together with the laws of the universe, we could predict with perfect accuracy the complete cosmic history.[4] Consequently, every choice is the result of the operation of the laws upon the world-states, set in motion ages past. If so, CON suggests, choices are unfree.

---

[4] Using Hempel's covering law model, let C_1, C_2, C_3 ... C_n be initial conditions, L_1, L_2, L_3 ... L_n the laws, and E_1, E_2, E_3 ... E_n the events. Given C_i and L_j, we can explain E_k, deductive-nomologically. C_i, L_j, and E_k are amenable to deductive inference, thus also to prediction.

Determinism is not unanimously held, since quantum mechanics suggests subatomic events are genuinely-metaphysically random. A minority of pessimists holds out the hope that the randomness is merely epistemic,[5] while most accept metaphysical micro-indeterminism, but hold it incapable of rising to the macro-level of choices. Hard determinists reject autonomy because pre-natal conditions determine each choice, whereas soft determinists deny this undermines self-control. "Hard" and "soft" may mislead; hard-nosed scientists may be soft determinists, for all determinists affirm the same causal/nomological history of choice – that all cosmic reruns are identical. Let us contrast logical/factual claims. A logical claim is that determinism is incompatible with autonomy. The hard determinist claims this and the factual claim, determinism. These claims imply the falsity of autonomy, as seen in the *modus ponens*:

HD1. If determinism is true, then freedom is false.
HD2. Determinism is true.
HD3. Thus, freedom is false.

The libertarian uses HD1 ("L1"), but, by contrast, makes a very similar *modus tollens*:

L1. If determinism is true, then freedom is false
L2. Freedom is true.
L3. Thus, determinism is false.

Hard determinism is thus both the logical thesis of incompatibilism and the factual/metaphysical thesis of determinism; libertarianism agrees about incompatibility, but affirms autonomy. A hard determinist, thus, is committed to logical and metaphysical claims, but also epistemic ones, to be spelled out shortly. Similar observations hold for libertarians and soft determinists.

The debate centers on the logical claims, but sometimes metaphysical and epistemic

---

[5] On the difference between ontological/epistemic readings of quantum mechanics, see Hodgson (2002) and Bishop (2002), especially pp. 116-118.

factors enter, and complicate matters. It would be fallacious, say, to infer incompatibilism from factual claims from psychology. Another fallacy is to reject libertarianism because it is improbable for neural firings to be influenced by randomly-emitted photons. Suppose no randomly-emitted photons ever influence neural firings, and suppose no cannibal indigenous to New Guinea ever earns a Ph.D. in philosophy from Oxford and becomes a professor. Both the 'cannibal philosophy professor' and 'quantum neural choice' doctrines are empirically possible, though improbable. Thus, the *empirical improbability* of quantum contracausality has no bearing on its *logical consistency*.

A "gnostic" has, and an "agnostic" lacks,[6] a pro- or con-attitude (true, false, or unknown) towards determinism, freedom, or their compatibility, computation of which yields 27 (3^3) possibilities: nine pairs of computations (3^2) for true, false, and unknown concatenations of freedom and determinism; add a factor of three for whether, given each of those nine positions, one is compatibilist, incompatibilist, or agnostic about compatibility, which yields 27 (3^3) views.

Logical and theological fatalism generate two more sets of 27 positions. Logical fatalists think bivalence entails nonautonomy; theological fatalists think God's foreknowledge ("theological precognitivism") entails it. Most theists are "soft theological precognitivists", analogous to soft determinists, affirming both doctrines; similarly, "soft bivalence" characterizes most who affirm bivalence. There may also be combinations of each of the three sets of 27 views, e.g., Richard Taylor (1983) combines hard determinism with logical and theological fatalism, and also degrees of belief.

A Procrustean approach might ignore the *heterogeneity* and *ubiquity* of autonomous

---

[6] There may be degrees of a/gnosticism. See Mele (1995, 2002) on agnosticism.

experience, *degrees* of belief, and *variations* in meaning. The *core notions* of the subject compete, e.g., causalist/contracausalist views of *possible* or *can*. Russell's close-, mid-, and horizon-level perspectives (2002) reveal three dimensions of autonomy that may lack a common feature. It is important to identify and remain sensitive to such complications in attempting to form the most coherent view. In introducing the issues, defining terms, and outlining positions on freedom, I have sought to remain loyal to these dialectical imperatives.

The major argument of the next chapter is that CON and other forms of pessimism have failed to remain loyal to these dialectical imperatives, and in so doing have given a false air of authority to fallacious ideas like actualism. The dethroning of these erroneous ideas will remove the false air of illegitimacy that has attached to ideas like conditional freedom and PAP, and will render optimism plausible in advance of positive arguments. I supply positive arguments in later chapters sufficient to ground a more robust theory of autonomy.

Fischer and other semi-compatibilists (e.g., F&R, 1998) agree with hard determinists (e.g., van Inwagen, 1975) and other pessimists (e.g., Pereboom, G. Strawson, Honderich, Smilansky, all 2002), and libertarians (e.g., Kane, 2002b) and other non-naturalist optimists (e.g., O'Connor, Ginet, both 2002), in holding that CON's *prenatal* determinism defeats *ultimate* control over who and what we become, choose, value, will, and do. That being so, Frankfurt (1971) and other soft determinists and compatibilists (e.g., Wolf, 1990, Dennett, 1984) have rejected PAP and the conditional theory of freedom, since CON rules out alternate possibilities and permits only a single past and future.

Instead, they (Frankfurt, Wolf, Fischer, F&R) opt for an account in which Agent is responsible *iff* he is in some sort of optimal internal relation to his will, values, character, choices, and/or actions in their actual-sequence unfolding, typically some sort of "mesh" between higher-/lower-order elements, which actual-sequence character would render them

consistent with determinism. These mesh or hierarchical theories are typically internalist accounts that eschew causal, historical, and/or counterfactual conditions, as these threaten to invoke CON's prenatal consequences. As such, they focus on what Russell (2002) would call mid-/close-range relations between Agent and his will or his will and actions. These accounts avoid horizon-level relations, for it is taken for granted that compatibilists cannot adequately dispose of horizon-level skepticism about prenatal control over the sort of character one is determined to develop. Incompatibilists (e.g., Pereboom, G. Strawson, Honderich, Smilansky, *id.*) have tended recently to unite on the common ground that mid-/close-range compatibilist senses of freedom and responsibility are plausible as they stand, but ungrounded at the horizon-level, since the sort of control one exercises in close-/mid-range domains is *ultimately* prenatally determined. The view from the prenatal horizon reveals no alternatives or control over who we become or what we choose, ultimately.

The latest stage in the dialectic involves a stalemate between "ultimatists", who agree that CON-ultimacy worries defeat all forms of autonomy, and "non-ultimatists", who allege that ultimacy worries do not defeat close-/mid-range control. But little argument is advanced by my fellow non-ultimatists against CON-ultimacy worries. Instead, focus on close- /mid-range matters is assumed to reveal their intuitive plausibility. Fischer (2002) says those who fail to get the intuition from Frankfurt-cases *that Agent is responsible when Intervener doesn't intervene* never will get it. I disagree, but postpone my reasons until chapter 5.

Giving causal/functional specifications for easy-autonomy would be insufficient for someone convinced that the causal/functional autonomy features it identifies lose legitimacy to prenatal-ultimacy worries. For then the putative autonomous agent is akin to a mere thermostat-/weathervane-type mechanism for which its every state is a mere function of ancient conditions/laws over which it lacks a say, which are not "up to" it, ultimately.

Thus, I argue against CON-ultimacy conclusions in chapter 2, which Archimedean point provides the leverage needed to dislodge the footing of the entire panel of debaters mentioned in this concluding section, and justifies my causal/counterfactual analysis against soft determinist claims that causal/counterfactual/historical conditions are irrelevant to freedom/responsibility. So, since pessimists of late adopt the stance that close-/mid-range models are coherent per se but vulnerable to horizon/ultimacy worries, once the latter are removed, this camp has reason to join ours. Similarly, once CON-ultimacy worries lose their sting, non-naturalist optimists no longer need embrace non-natural accounts of autonomy to be optimists, so they, too, have reason to join our camp. Finally, once determinism is no longer a threat to the self-control soft determinists and other compatibilists once embraced as intuitive forms of autonomy, they no longer need to reject PAP and the conditional theory of freedom, so they, too, have reason to join our camp. Thus, with my anti-ultimacy arguments in place, all forms of pessimism and non-naturalist optimism lose footing, as do all forms of soft determinism and compatibilism that reject PAP and the conditional theory; they are left standing without grounds for their views, and are invited to join our ranks. The theory they are invited to embrace may be more robust than previous attempts along similar lines. For hierarchical, actual-sequence, and other internalist accounts may now be enriched by PAP, conditionalist, counterfactual-sequence and externalist, e.g., causal and historical, criteria, paving the way for easy-autonomy. Indeed, once CON is dislodged, that all hard-problem-motivated accounts are otiose is overdetermined by this and the easy-approach.

Dennett claimed (1984) there will never be 'a' solution to the freedom problem, because there are many such problems. I collect them all, link them to actualism, dislodge it, and thereby render our daily conception plausible. That is my simple – albeit highly ambitious – plan. It is supported indirectly by a claim Wolf made (1990) that all standard

conceptions of autonomy are incompatibilist, since she thinks they require indeterminism or non-naturalism. Wolf is mostly right, and it is CON that is behind this intuition, the same intuition that led F&R (1998) to leave autonomy out of their semi-compatibilism, though that is a hair's breadth away from an autonomy account, unbeknownst to them. What is unique about my account is that it is a form of *compatibilist autonomism*, what Wolf thinks is oxymoronic. This is a mixed blessing. If the notion is oxymoronic under the current paradigm, but a plausible model of that notion is instantiated here, then that is something worth knowing. But if the notion is canonically-viewed as oxymoronic, then my work is cut out for me, but the expected harvest is well worth the toil.

Chapter Two: A Critique of Pessimism

"Past-Fixity", "Present-Fixity", "Future-Fixity", "Laws-Fixity", and the "Uniformity" of *same-world/same-history-and-laws* are *axioms* of determinism, together, "Fixity". Fixity drives van Inwagen's (1975) CON, a classic expression of incompatibilism. I argue in this chapter that CON commits several errors in connection with Fixity, especially equivocations, and Lehrer's critique of conditionalism commits similar equivocations, as does fatalism.

The contracausalist thinks Shelley could have chosen ~A contrary to M1 (a subset of C) that caused A. The counterfactualist thinks she could have chosen ~A *only* if ~M1 (thus ~C) held. C should include *only* circumstances that contribute to A. If *all* pre-A facts/laws are relevant, they *globally* constitute C, so restriction *to relevance* eliminates nothing. But if causation is *local*, the restriction preserves truth. But it is implausible that every fact/law is relevant to every other one. Shelley's thirst/water-beliefs are the relevant causal conditions for A, not Caesar's crossing the Rubicon.

Taylor and Dennett (2002) think causation talk is pragmatic, limited to local causes. The Big Bang didn't *cause* the Twin Towers' collapse. The *hard determinist* may reply that whenever the cosmic film reaches C, A results, *ad infinitum*, so Shelley doesn't control whether A occurs. But this ignores that the causal determinants in C are *precisely* her desires to do A under M1 that comprise C, *locally* construed. A causal dilemma arises:

CD1. Either C is construed globally or locally to cause A.
CD2. If C is construed globally, it misrepresents irrelevant factors as causally relevant to A, but *a causal factor F is relevant to E only if a change in F would bring about a change in E*, and not all factors in globally inclusive C satisfy this causal criterion.[7]
CD3. If C is construed locally, it involves M1, Agent's beliefs and desires, in just

---

[7] See Lewis (2000) for a defense of the italics in CD2. A similar notion is widely accepted by statisticians to the effect that variable F is causally relevant to variable E *iff* variation in F "explains some of the variance" in E, i.e., holding F constant reduces variation in E.

the right way, such that they satisfy the causal criterion (see CD2 above).

CD4. Thus, either way, C's causing A does not undermine autonomy.

The global configuration of effects is an *indirect* function of the global configuration of causes, since each cause/effect pair has only *its* linkage. C_1, C_2, ..., C_n cause actions A_1, A_2, ..., A_n, and laws L_1, L_2, ..., L_n relate them *as pairs, not summed.* It is an agglomerative fallacy to sum local causal relations globally. *Though state descriptions entail each other under determinism, world-states do not cause each other, so there are no world-state laws.*

There are other reasons to reject global causation. Despite the bluff that quantum indeterminacy 'cancels out' at the macro-level (Honderich, 1993), indeterminacy *anywhere* renders *deterministic* state-description laws *impossible.* Physicalist reductionism denies the causal efficacy of mental states, as if identifying the locus of causation at such special-science levels of description constitutes a category error. But if true causal explanations are only atomic and there are no licit special-science-level attributions of causality or laws governing macro-level objects such as mouse-traps (Schiffer, 1991) or bowling balls, then the implicit principle is that *it is illicit to use causal/nomological language at the macro-level.* Since state descriptions are macro-agglomerations greater than bowling balls, they cannot be the relata of laws. *If there are no mouse-trap or bowling ball laws, then there are no state-description laws.*

Consider the alleged need to reduce folk psychological generalizations to physics.[8] As Schiffer argues, even if there are ways in which contingent generalizations about how we reason *may* be reduced to more basic levels (to complete their *ceteris paribus* clauses),

> it's by no means obvious that ... a reduction to fundamental physics, or even to neuroscience, would be needed to provide that explanation (1990, p. 171).

---

[8] On such reductionisms, see Kim (1979, 1984, 1989, 1992, 1999, 2000); *cf.* Schiffer (1990, 1991), G. Strawson (1997); on reductionist determinism, see Honderich (1993, 2002).

Schiffer supports this claim by reference to

> a well-known experiment ... (Wason and Johnson-Laird 1972), which illustrates nicely the sort of contingent generalizations that psychologists feel a need to explain.
>
> You are presented with four cards, each having a letter on one side and a number on the other. The up sides show the following characters:
>
> E K 4 7
>
> Which cards need to be turned over to test the following conditional? If a card has a vowel on one side, then it has an even number on the other side. 'E' and '7' is of course the correct response, but Wason discovered that a common response is to pick only the card with the vowel. (Schiffer, *id.*, pp. 171-2).

These results seem to show that content has something to do with reasoning, not just syntax.

> But suppose some joker tried to explain the generalization that people tend to overlook the '7' ... by deriving it from quantum mechanics.... Even if this could be done (which I seriously doubt) it wouldn't yield an explanation that any psychologist cares about; the [psychologists'] questions ... would still be unanswered. More pertinent, though, is the observation that it may be doubted whether the reduction to fundamental physics is in any sense *needed* to explain the generalization. (Schiffer, *id.*, p.172; emphasis in original.)

There is no need for mouse-trap *laws* or reduction of their related law-like generalizations.

The reductionist claims explanation is micro-level, but when it's pointed out that the micro-level is *indeterministic*, he uses a "micro-macro prestidigitation" to shift to the special-science claim that psychological determinism holds. Laws that *do* cover global distributions of particles, e.g., of thermodynamics, are not about *local* causal linkages, not about C. Thus, "state description" and other global terms cannot play the right roles in a *causal* account of action.

Slote expresses similar intuitions about necessity (1982), and despite technical objections, he is correct; *cf.* Kapitan (2002). Causal necessity is selective and cannot be agglomerated at the level of state descriptions that entail each other under determinism. What happens in a rice paddy at T_x in China, part of world-state P_x, does not influence what happens in Shelley's home at T_y, part of world-state P_y. It *could* influence her, but not in the global sense assumed here, so building what happens in the rice field into the cause of her

behavior is gratuitous and misleading. *World-states do not cause local states.* Formally, the occurrence/non-occurrence of event E is a function f of variables $v_1, v_2, \ldots, v_n$. That is, $E=f(v_1, v_2, \ldots, v_n)$. Hence, the values of any variables other than the v_i are irrelevant; thus, $\sim v_i$ (a rice-paddy event in China) is irrelevant to f, thus to E (Shelley's drinking).

Focus on global reruns would distort the explanatory narratives that inform our *local* biographies. Only the film with the narrative structured around the formation of Agent's character and experiences captures the relevant features of his life; a good biography places its subject within an environmentally-sensitive rational structure. With advances in the special sciences that explain human behavior, future biographies will contain more extra-agential facts, but these will be limited to the vicinity of the biographed person's life.

Seeing such a film of our own lives invokes an initially-fatalistic dread, but this proves no more than discoveries about the neural basis of this or that surface-level behavior. Richard Taylor (1983) describes a fictitious "Osmo" who discovers a book about his life, and his attempts to avert his own death cause it. Goldman (1970) and others deploy "book-of-life" scenarios, but these Oedipus-style fatalisms require a *deux ex machina*, and so are not useful in resolving puzzles of self-prediction, as in Newcomb's paradox. For *deliberation whether to do A* cannot be maintained in the face of *knowledge that one will do A.*[9]

The *global* rerun view reduces choice to *discovery* of what one will do, no matter what one may try to prevent that from happening.[10] The phenomenology is that Shelley's

---

[9] Goldman (1970) resists an incompatibilist intuition: If the book of all facts about my future were presented to me, I could easily falsify it. Determinism initially appears incompatible with this ability, but my chapter 1 arm-raising analysis suggests otherwise.

[10] Buddhist "enlightenment" involves consciousness of the flow of impersonal influences through the mind-body in the absence of an integrated self. Here, the dreaded image is embraced as the epiphanous peak of a lifetime of soteriological inquiry. See G. Strawson (1986, pp. 117-20) for a Zen-like version of hard determinism.

choosing is a *doing*, something she *makes happen* by an *act of will* – not a *passive discovery* of what determinism entails for her as a *mere container of effects*. It is preferable that C causes A, for if C yields what should be an effective desire for A, water, but Shelley does B, shoves salt in her mouth, this would be pathology, not autonomy. This is not "sweet lemons" (Elster, 1983), but the intuition that Agent cannot claim authorship of A if A is random, not caused by C. Autonomy *requires* determinism, or else pathology, B, would be "freedom". Determinists agree on the *necessity* of identical reruns under Uniformity for authorship.

To circumvent the optimist's focus on such sufficiently explanatory features as those of local C, the pessimist notes that C was caused by earlier C-1, in turn by C-2, C-3, back to prenatal conditions C-n. If her life is determined by events over which she had no control, her autonomy is illusory. Pessimists insist a narrow focus on Shelley's segment of the cosmic-history clip misses the larger narrative.

Since prenatal C-n determined A, Shelley couldn't *really* have brought about ~A, so she is unfree. In a sense, CON says here that determinism implies there is no autonomy under the *contracausal* conception that involves ~A, and thus *indeterminism*. But *determinism's incompatibility with indeterminism* is tautologous. The *counterfactual* conception remains intact. CON thus equivocates about the "freedom" it targets. Either CON targets deterministic or indeterministic models. CON mainly assumes an indeterministic model, so it doesn't directly target soft determinism, but libertarians are already incompatibilists.

CON is supposed to *support*, not *express*, that determinism is incompatible with

freedom.[11] CON assumes freedom is indeterministic and then shows that this indeterministic freedom is incompatible with determinism – but *of course indeterminism is incompatible with determinism*! CON presents itself as if it has refuted deterministic freedom.

Though abstracta cannot impel actions, CON suggests state descriptions and laws render Shelley unable to refrain from drinking that water. Tautological egoism holds *action* self-interested because *intentional* by definition, what Agent thus *wants*, hence *Agent interest*, and so self-interested. But this tautology doesn't prevent the *content* – the aim – of a motive from being other-regarding as well. So, too, abstracta about determinism don't prevent agents' *control* of specific actions.[12] If Shelley cannot refrain from drinking water, it must be due to a brain or behavioral disorder, not abstract puzzles about modals.

Similarly, though Zeno is best defeated with calculus, nobody thinks logic immobilizes, much less proves motion nonexistent. Just as motion is ubiquitous, so, too, autonomous actions are ubiquitous. Presumptive favor goes to motionists and autonomists; the evidence massively confirms both. If we cannot move or control our actions, it is not due to abstracta, but to some extra-logical pathology. Belief in motion and autonomy are justified, by virtue of their ubiquity alone, in laying claim to presumptive favor in debates about their legitimacy. Just as Moore famously 'proved' there is a world by raising his hand, or at least gave a demonstrative reminder about *the evidential priority owed to our common experience*, so also a person can 'prove' his autonomy by raising his hand. One hand movement, three proofs!

---

[11] Levin (2003b) says unfreedom is supposed to be an interesting *consequence* attendant upon considerations about determinism, not part of the thesis itself. That is, it is arguable – indeed, *the* argument – whether to interpret determinism as soft or hard determinism.

[12] Though it is doubtful hard determinists advocate it *as such*, the opposite view is an instance of what Whitehead called the "fallacy of misplaced concreteness" (Landesman, 2003).

In his seminal paper, van Inwagen (1975)[13] uses "P" to refer to the state description at time T when judge J refrains from raising his hand (to signify a judicial decision), "P_0" to refer to whatever state description is true at prenatal time T_0; and "L" to refer to the conjunction of all the laws of nature. So,

(1) If determinism is true, then the conjunction of P_0&L entails P.
(2) If J had raised his hand at T, then P would be false.
(3) If (2) is true, then if J could have raised his hand at T, J could have rendered P false.
(4) If J could have rendered P false, and if the conjunction of P_0&L entails P, then J could have rendered the conjunction of P_0&L false.
(5) If J could have rendered the conjunction of P_0&L false, then J could have rendered L false.
(6) J could not have rendered L false.
(7) If determinism is true, J could not have raised his hand at T. (pp. 52-3)

Van Inwagen says the conditionalist thinks sentences of type (8) *mean* sentences of type (9),

(8) S could have done X.
(9) If S had chosen to do X, S would have done X. (p. 57)

and challenges the conditionalist with the following dilemma. If we so substitute each sentence in CON, either it will remain true or, if false, that constitutes a *reductio* against the substitution (p. 58). Van Inwagen leaves this task to the reader. We focus on (1)-(7) first.

P_0 has only *global* content about T_0's state description, whereas P has *local* content about J's action. But the *global* view distorts vision in favor of hard over soft determinism. Also, J's refraining from raising his hand involves an omission, rather than a positive action, which counts as a judicial decision. In other contexts, omissions typically lack this feature.[14] Later CONs lack this feature, but it is important to note its role in original CON.

---

[13] Page numbers refer to Watson (1982). Recent supporters are G. Strawson, Pereboom, Honderich, Smilansky, Fischer (Kane, 2002a); earlier, Hospers, D'Holbach, Lucretius.

[14] Analogous to *locutionary actions* as utterances of marital "I do", this *nonlocutionary omission* is an expressive gesture without movement and has performative force as a convention-generating action. These complications may warp intuition.

Some objections. (1) trades in *local causal* matters for *global entailment* ones. These substitutions distance intuitions. (2) is true, but its antecedent uses counterfactuals in a way that has agents *do* other than they do, which suggests counterfactuals need to be actualized. (3) is a *non-sequitur*, since (2) does not entail (3)'s consequent, and the truth of a counterfactual does not entail ability to falsify the factual statement it runs counter to. Though (2) and the conjunctive antecedent of (3) ("ca3") (by exportation) are true, the consequent of (3) ("c3") doesn't follow.

(ca3) (2) is true and J could have raised his hand at T.
(c3) J could have rendered P false.

For (2) is not a factual claim that J *did* otherwise, and (ca3)'s latter conjunct is merely an ability statement, whereas (c3) asserts J's ability to actually falsify a fact about what J is stipulated to have already done – that's impossible *tout court*, barring time travel.

(4) and (5) inherit the illogic of (3), and (5) assumes ability to do otherwise implies ability to violate Laws-Fixity, for it goes without saying that we cannot violate Past-Fixity; time travel is conceivable, but not as a basis for autonomy. But rather than imply a power to change the past, Uniformity requires that counterfactual reasoning be varied across the board, i.e., the counterfactual past must differ if supposing a counterfactual present. CON implies the only possibility is that *able to do otherwise* entails *able to violate the law*. Since putative 'miracles',[15] strictly, cannot be more than merely highly-unexpected phenomena unless they violate a law of nature, such miracles are incredible, so van Inwagen asserts in (6) that we lack this ability.

Lewis says van Inwagen's "render false" notion is ambiguous. In the *strong* sense, J

---

[15] Miracles are impossible on that view that sees laws as exceptionless generalizations, in which case any putative 'exception' would automatically invalidate any previously-considered law it was seen as an exception to, and convert it into an 'almost law'.

is able to do something such that a proposition is thereby *made* false by his act, but in the *weak* sense, J is able to do something such that a proposition is thereby *shown* to be false (1986b, p. 293). On the weak sense, (6) is false, according to Lewis (1986b, p. 297), for there may be "divergent miracles" if at T_{-1} God replaces one set of laws with another, in which case J's *doing otherwise* at T merely *shows* the law false (1986b, pp. 294-95).

Call the different laws involved at T_{-1} "L*". Though a special *attributive* reading of "law" as *whatever* is an exceptionless law-like generalization requires all possible exceptions to revoke the status of a putative law, "law" ought not to be so defined that it is linguistically impossible to *describe* worlds in which God holds fixed, say, their initial conditions, but varies their laws. Thus, if the laws *were* changed by God, such that J *does* raise his hand, then J does something such that the *previous* set of laws L is *shown* false thereby.

This weak/strong distinction can be applied to CON another way. Only the *contracausalist* claims that J could have brought about ~P with the same history/law that entails P. The counterfactualist claims J could have done otherwise *if J wanted to*, which assumes a different past/laws $\sim(P_0\&L)$; J doesn't *do* otherwise in any same-world reruns. If J *had* wanted to do otherwise ~P, Uniformity implies things would have been different back to T_0, and, under this counterfactual, P_0 would be false. There is no problem with $P\supset(P_0\&L)$ and $\sim P\supset\sim(P_0\&L)$.

Call conditions acting on J at T_{-1} "C_J"; call Shelley's "C_S". For counterfactualists, if J *had* wanted otherwise, J *would have been* under $\sim C_J$, entailed by $\sim(P_0\&L)$. *Contracausality* requires the strong "render false" where C_J produces *different* reruns. Contracausalists say *~P can follow* $P_0\&L$, but since P_0 is set at T_0, P_0 remains fixed, so L is falsified at T. Contracausality doesn't violate Past-Fixity, therefore, but Laws/Present Fixities.

Counterfactualists do not claim J can select from alternatives Pv~P. *All* determinists

accept Fixity, so CON cannot target counterfactualism, only contracausality. But *everyone already sees* determinism as incompatible with indeterminism, so a CON dilemma arises:

COND1.	Either CON targets soft determinism or libertarianism/indeterminism.
COND2.	If CON targets soft determinism, it is irrelevant, since soft determinism is an axiom-observing, merely counterfactual model of freedom.
COND3.	If CON targets libertarianism, it is redundant, since libertarians already accept indeterminism and its obvious incompatibility with determinism.
COND4.	Thus, either way, CON is idle.

J's *ability* does not require *strong/contracausal* possibility ~P, only that J *would have* done otherwise *if he wanted to – weak/counterfactual* possibility. Even given C_J, J still *could* have chosen otherwise *if* he had wanted to, if $\sim C_J$, but C_J holds. *Weak/counterfactual* ability doesn't render ~P; P holds.

That it is determined that J didn't want to do otherwise doesn't rule out that J could have so wanted, *had he wanted to* want otherwise. It's all determined, but J's abilities remain, without becoming law-breakers/fact-falsifiers. J cannot alter P, P_0&L, or $P \supset (P_0 \& L)$. *Contracausalists* alone bear the burden of facing these antinomological implications.

PAP has come under attack from Frankfurt (1969). Given that Intervener does not actually interfere in stipulated cases where Agent wants what he wants, Agent acts on his own desires, so he is still responsible for what he does, despite the fact that Intervener bars his access to alternate possibilities. PAP may be read two ways. *Strong PAP* views alternate possibilities *contracausally*; *weak PAP* views them *counterfactually*. Strong possibilities involve *agent-access* to divergent futures in choice, thus non-uniform reruns in the *same* world and violate Fixity. Weak possibilities only involve paths that *would* have been open *had* different causal conditions led to choice, thus uniform reruns in the same world; these only *appear* to allow non-uniform reruns if *wrongly* viewed *across* worlds, but not *within*

them. So, in the next world in which $\sim C_J$, determinism entails ~P – there are no axiom-violating possibilities *within* worlds, so weak possibilities are determinism-friendly. Fixity holds in all counterfactual worlds; these are *not agent-accessible.*

In recent discussions (e.g., Dennett, 2003; Levin 2003), disagreement about the need for PAP among compatibilists suggests the issues have been miscast from the outset. Some reject PAP (Frankfurt, 1969, and Dennett, 1984), but some don't (Levin 2003), and the split among them can be mended by my bifurcation. All determinists reject strong possibilities, and since strong PAP requires them, they must reject it. Hard determinists insist freedom requires strong possibilities and thus is impossible. But all determinists must accept weak possibilities, since the opposite doctrine – that only what actually happens is possible, "actualism" – is not an option for them, I will argue. Weak possibilities, thus weak PAP (see PAPW, below), are *not* determinism-unfriendly.

Bok (1998) distinguishes practical/theoretical possibilities. Any option Agent considers is one that *would* open a metaphysical path *were* Agent to choose it. Given Agent's power to determine a metaphysical option by selecting it, all considered options are practical possibilities for Agent, though only Agent's choice is theoretically possible. Bok's practical possibilities are our weak possibilities, for each option in the vicinity of being chosen represents a weak possibility in a nearby alternate world which would be available, were it selected. For another to be selected, however, a different past had to hold.

The intuition behind weak/practical possibilities is captured in "PAPW", a PAP-like principle highlighting the word "would". Here is a sketch of PAPW, which has two parts:

PAPW1: Agent *voluntarily* does A *iff* he *would* have done A *had* he been able to do otherwise;

PAPW2: Agent freely does A *iff* he does A voluntarily and *would* have done other than A *had* he wanted to.

PAP requires agent-access to alternate possibilities; PAPW does not. PAP focuses on *normative* responsibility; PAPW focuses on *causal* responsibility, *authorship*.

PAPW1 explains our willingness to hold agents responsible though they lack alternatives. For they would have engaged in these actions even under strongly-construed alternatives; PAPW2 captures both Bok's intuitions and weak conditional possibilities.

CON's (3) assumes *able to do otherwise* means *able to actualize either of Pv~P under* C_J, which violates Present-Fixity. This is strong contracausal ability/possibility. *If we must imagine ~P, CON cannot validly require us to hold* P_0*&L fixed* under Uniformity. CON equivocates by treating P *variably* but P_0&L *rigidly*, asks us to run our P-*variable* intuitions against a P_0&L-*rigid* past, and then counts our failure at this task as a defeat. But it begs the question to suppose ~P, but not to suppose ~(P_0&L), to suppose contracausality.

If we treat P and P_0&L symmetrically, we see that counterfactual ability only entails that if J *had* done otherwise, this would *show*, in Lewis's weak sense, that J *had* a different history ~C_J, in which *he wanted otherwise*, which requires different antecedents back to T_0, in which case P_0&L would be *shown* false. The weak sense of "render false" entails that (5), (6), and (7) fail. (5) assumes that J cannot render P_0 false. But if we assume ~P while P⊃(P_0&L), but we are not permitting the asymmetry whereby P is but P_0 is not varied, we cannot justify the assumption that P_0 cannot be falsified.

The problem's root is *the very assumption* that *ability to do otherwise* entails *ability to bring about ~P*. For if J *is able* to do otherwise at T, then that full-range motion-capacity would be instantiated in J's skeleto-muscular configuration and so must constitute part of P's content. So, if J has the ability, *P remains true*, for P is *constituted* – not *contradicted* – by the causal functional powers J possesses at T. Truth-conditions for P are that J's arm is not broken and his brain is functioning normally, so if an intention/neural-signal was formed to

raise the arm, it would go up. To *count* J's *not* raising his arm *as a judicial decision* requires viewing it as a *non-catatonic omission*, otherwise the case could be appealed on the ground that *J's arm couldn't go up.* Thus, it is a *non-sequitur* to assume that *J's doing A while being able to do ~A* (both described in P) involves *J's being able, while P, to bring it about that ~P.* One way to avoid this is to go atomistically/actualistically reductive, so P is written in terms of particles/trajectories, but we have rejected micro-macro prestidigitation; besides, enough linked-trajectories constitute ability (e.g., continence)[16] or disposition.

That holding P variable but P_0&L fixed is fallacious is unnoticed due to Past-Fixity, which is *a priori* relative to determinism. There is a crucial difference, however, between being able to violate Past-Fixity and being able to imagine a different past, given a different present.[17] Philosophers trying to get around Past-Fixity fail to notice that *supposing ~P violates Present-Fixity.* If we alter one, Uniformity *requires* we alter the others. CON's equivocation *smuggles in contracausality* in supposing ~P *in the P world.*

J's ability never falsifies P, so J does not falsify P_0&L. (5) erroneously holds P_0 fixed but treats P as variable, which equivocation suggests J could have rendered ~L. Now, only the stronger "render false" *makes* the past be actually, rather than counterfactually, changed. The weaker "show false" does not *violate* Past-Fixity.

Though ~P is *not required*, when asked to suppose ~P, it is intuitive to think the past

---

[16] This power manifests differently, e.g., when C_{J1}, C_{J2}, ... C_{Jn} hold, P_1, P_2, ... P_n result, respectively. This is confirmed through countless observations in ubiquitous experience, where different PAPW-satisfying agent-intentions result in concordantly different actions.

[17] Past-Fixity may be *synthetic a priori*, since its necessity rests on unidirectional temporal succession. It is so entrenched that reasoning that *seems* to challenge it may be outlawed without a hearing. At the end of *Euthyphro*, Socrates asks *why the gods encourage piety though it doesn't benefit them*, and Euthyphro says "because it pleases them". Socrates rejects this because it is the same outlawed *phrase* used to *define* piety. This is a fallacy, for homonymy is not substitutable across divergent contexts. I call this the "outlaw fallacy", as if a phrase, once designated illicit in a context, remains 'outlawed' in every context. Those who invoke Past-Fixity without noticing its mere homonymy with P_0 commit the outlaw fallacy.

must have been what was different, rather than the laws, once we insist on compliance with Uniformity. Since the smallest changes in our conceptual system are always preferable, it is preferable to imagine some *earlier move* in billiards varied, than to imagine *the rules* varied. To hold constant the balls' movements in a game's history while varying the rules, so that the loser would have won (e.g., the 'winner' is the racer who comes in *last*), defeats the purpose of counterfactual reasoning. When someone claims he could have won, he means under the same rules, with reasonable variation in facts consistent with his abilities.[18] The nearest possibility is one which preserves laws (Lewis, 1979).

Given the same laws, scientists imagine divergent initial conditions in a system, producing a variety of different system-evolutions, *de rigueur*. The experimental method's *repeatability* criterion, scientific laws' *abilities to support counterfactuals*, and scientific knowledge *in general* support the notion of counterfactuals within deterministic contexts without supposing changed laws, so there is little reason to think differently here. "Had the billiard ball been struck at a different angle, it would have rolled in a different direction" is a paradigm of Newtonian determinism. While conceivable, different laws are not what we mean when we say things could have turned out differently. Thus, (5) fails.

The usual place to look for change in a system is the initial conditions, but the laws *could have been* otherwise. Thus, if J *had* done otherwise, thereby merely *showing* P false, thereby *showing* P_0&L false, L *would* be false *if* initial conditions at T_0 were held fixed but indeterministic. So, (6) fails. Since (5) and (6) fail, (7) does too. For what was supposed to entail (7) was an inference which, if simplified, on the stronger sense ("SS") of "render false", is "SS Nomic CON":

---

[18] Taylor and Dennett (2002) discuss how to use nearest-world counterfactuals to analyze abilities.

SS1. If J can do otherwise, then J can make it the case that ~L (can violate Laws-Fixity)
SS2. J cannot make it the case that ~L (cannot violate Laws-Fixity)
SS3. Thus, J cannot do otherwise. MT, SS1, SS2

SS1 is a summary of (1)-(5); SS2 restates (6), and SS3 restates (7). But SS1 is false.

On the weaker sense ("WS") of "render false", the claim that J can render ~P – violate Present-Fixity – doesn't arise, but the argument would be "WS Nomic CON":

WS1. If J can do otherwise, then J can do something that shows ~L.
WS2. J cannot do something that shows ~L.
WS3. Thus, J cannot do otherwise. MT, WS1, WS2

If we assume that J can counterfactually render ~P, we must treat "L" symmetrically; so, WS2 is false. When we make explicit the enthymematic rejection of $\sim P_0$ in original CON, CON fails. On the stronger "render false", the argument is "SS Historical CON":

SS4. If J can do otherwise, then J can change the past (can violate Past-Fixity).
SS5. J cannot change the past (cannot violate Past-Fixity).
SS6. Thus, J cannot do otherwise. MT, SS4, SS5

But SS4 is false, the same as SS1. The weaker argument is "WS Historical CON":

WS4. If J can do otherwise, then J can do something that shows a different past $\sim P_0$.
WS5. J cannot do something that shows a different past $\sim P_0$.
WS6. Thus, J cannot do otherwise. MT, WS4, WS5

Again, if we assume ~P, we must also assume $\sim P_0$, so WS5 is false.

If asked to suppose *P&~P*, we'd laugh if inability to do both in the same universe of discourse were seen as a dialectical failure. This would be a "2-card monte". In 3-card monte, target card is shown face up, added to two others, all are faced down and moved rapidly by dealer, and gambler's task is to identify target card afterwards. The difference between CON and 2-card monte is the number of factors, but CON's complexity hides the fallacious elements like 3-card monte's third card. CON assumes ability entails ~P, then focuses on what would be falsified if ~P, but which agents cannot falsify. P is *varied*, but

P_0&L is *fixed*. This equivocation treats same-world-linked propositions asymmetrically,[19] but Uniformity requires their *equal valence*.

Fixity requires that J never *does* otherwise. CON assumes reruns can differ in *only one part of a same-world series*. But Uniformity requires that if any same-world-W difference is assumed, adjustments must extend throughout W. Thus, to vary P but not P_0&L violates the supposition that frames the debate – Fixity. J only could have *done* otherwise *if* previous segments in W differed. The ubiquitous ability to carry out different actions under *different* intentions is Fixity-friendly. There are indefinitely many times in our lives when our wants change, and we do different things. Wants matter.

Counterfactual-supporting laws are inferred from *actual-sequence* evolutions in deterministic systems in which experimenter-agency – where experimenter's actions introduce catalysts into a system – is bracketed off. Scientists thus infer counterfactuals from deterministic systems by wiggling the variables (Taylor and Dennett, 2002) and bracketing off their own interventions (also presumed determined), without any Fixity-violations; we may likewise infer our counterfactual powers to alter our wills/actions by wiggling the variables in ourselves. Were global causal inferences licit *tout court*, *experimenter interventions* could not be bracketed off, local causal inferences would be illicit *tout court*, and science would collapse.

When we want to do different things and do them, PAPW2 is ubiquitously confirmed, but when we fail, we experience weakness of will; if chronically, we experience deeper

---

[19] That they are linked does not mean that they entail each other *per se*, absent laws, but *with* the laws, given that they are of the same world. Van Inwagen insists that we must construe state descriptions as not containing information in them about the past or the future, but only about what holds in the universe at an instant, for "in that case, determinism would be a mere tautology.... This amounts to saying that the 'laws of physics' clause ... does some work: whether determinism is true depends [on] the character of the laws of physics" (1975, p. 48).

privation of autonomous control over that dimension of our lives. PAPW2 is violated here though PAPW1 may be satisfied, but its violation constitutes indirect support for its validity, for its violation explains privation as PAPW2-failure.

Akratic behavior is *occasional* PAPW2-violation; compulsion involves *chronic* violation. Both types are fairly local in that *sufficiently autonomous* agents *usually* maintain a *background* of sufficiently PAPW-satisfying activities, such as eating when hungry. Some psycho-pathologies involve more global PAPW-violations, e.g., motor disturbances; these may be insufficiently autonomous. To the extent PAPW is satisfied, judgments about personhood, responsibility, and exculpation fit. PAPW-satisfaction captures the voluntary; that it correlates with judgments of responsibility constitutes reflective equilibrium.

Thus, the hard determinist cannot establish the epiphenomenal point that motives are causally impotent, since it is *precisely* the extensively-justified core-science tenet that one local causal-nomological state is sufficient for the next that constitutes the *content* of determinism, as the global causation objection makes clear. Conversely, doing different things under different wants makes determinism intuitive. The ubiquity of causally-effective linkage between volitions/actions constitutes *massive confirmation* for daily autonomy *qua* causal control of intentions/choices over actions and thus their counterfactual power as variables – when others are present, we do other things. Determinism is an abstraction from countless counterfactuals and is made personal by countless experiential/folk generalizations about human behavior.[20] Autonomy implies no Fixity-violating ability. In suggesting otherwise, CON commits the straw man fallacy.

---

[20] This phenomenological component of experience, Hume thought, grounds the sense of necessity in the notion of causation, and though he was a projectionist about it (1748, sec. 7), the point remains: What links choice *causally* with action is that ubiquitous experience confirms the association, and when this link is broken, there is a deep privation of agency that exculpates.

Let "P*" refer to the state description at time T when the 8-ball does not go into the corner pocket, and "P^*_0" refer to the state description for some pre-game time T_0. "L" has the meaning it has in CON. Our parallel CON with billiard balls is "BB CON":

BB1. If determinism is true, then the conjunction of P^*_0&L entails P*.
BB2. If the 8-ball had gone into the corner pocket at T, then P* would be false.
BB3. If BB2 is true, then if the 8-ball could have gone into the corner pocket at T, the 8-ball could have rendered P* false.
BB4. If the 8-ball could have rendered P* false, and if the conjunction of P^*_0&L entails P*, then the 8-ball could have rendered the conjunction of P^*_0&L false.
BB5. If the 8-ball could have rendered the conjunction of P^*_0&L false, then the 8-ball could have rendered L false.
BB6. The 8-ball could not have rendered L false.
BB7. If determinism is true, the 8-ball could not have gone into the corner pocket at T.

BB7 says that if the 8-ball didn't go into the corner pocket, it couldn't have, which betrays actualism. But "counterfactualism" underpins 'law' as what *supports counterfactuals* and makes determinism tenable at all. "Actualistic determinism" is oxymoronic.

Uniformity and experimental method presuppose that objects have nomic – dispositional/counterfactual – properties, so uniform results follow uniform experiments. Determinists cannot take a reductive view toward nomicity; *nomic* means *counterfactual supporting*. Determinism *just is* an abstraction from the nomological character of *all* causation, and so cannot be eliminated without throwing away the baby with the bath water.

Even if there were good reasons to avoid them, the same notions can be expressed by putting a new dress on them, e.g., by talk of propensities or potential evolutions of systems, and the new terms would still have to house the related notions *required by determinism.* Given the functional equivalence *under determinism* of such reductively-modified notions and the oxymoronic character of *anomic determinism*, since *maximally anomic* contradicts *maximally nomic*, pessimists do not have the dialectical luxury of appealing to anti-

modalism, reductionism, or any other form of actualism. Even if not, actualism is more dubious than the doctrine it seeks to negate, and so begs the question. Determinism requires counterfactuals; so, *had* an 8-ball been struck differently, it *would* have behaved differently. BB7 is contradictory and question-begging.

Notice the move from BB2 to BB3, from *supposing* P* to be false to this 'rendering' P* false. *Agents* can 'render', but billiard balls cannot 'render' anything. I ask the reader: *Why use "renders" rather than "entails"?* "Render" cloaks the strong/weak equivocation, attributes inflated powers to agents, and facilitates the intuition that freedom requires impossible abilities. If "entails" were used instead of "renders", there would be no such implication, for *propositions entail, not agents*. If asked to assume a factual statement, then to entertain its denial, the mere fact that its denial entails its falsity – or that of propositions which entail it – would not lead anyone to think that anything peculiar was going on; that's why there are no 2-card monte hustlers out there. Thus, "renders" distorts judgment, when P_0&L, held sacrosanct by misconstrued Fixity, appears 'rendered' false by J.

There is nothing worrisome about how an 8-ball that didn't go into a corner pocket (P*) might have. To suppose that it did (~P*) is to suppose a counterfactual that would make P* false. Counterfactuals *entail* the falsity of facts they *counter*, without *rendering* Fixity-violations. Both P* and ~P* are suppositions in the same discussion, each has a different world-history by Uniformity, and each entails the falsity of the other. Thus, we begin with a supposition and then face it against its negation in such a way that the variable one is expected to defeat the fixed one *somehow*. But because only one is held fixed illegitimately, our inability to do the illogical is seen as a failure. This impossible task and its dyslogia are cloaked by equivocation. Both suppositions are of equal logical force; so long as the truth-values of their entailments are both held fixed or both varied, they are innocuous.

The solution, which the weak "renders false" permits, is to eliminate equivocation by holding both terms equally. If the ball doesn't go into the corner pocket, it has one world-history; if it does, it has another. Shifting between suppositions – P*/~P* – doesn't falsify anything; neither supposition has priority, and both require the other's falsity and that of the other's antecedents, but so what? Nothing about this entails Fixity-violating ability.

Newer, "master-CON", reifies these errors with greater formal abstraction, "α" and "β" rules, and "□" and "N" operators,

α □p ∴ Np
β Np, N (p ⊃ q) ∴ Nq

where "□" expresses logical necessity and "Np" expresses "the universal unavoidability of p" (Kapitan, 2002) or "p and no one has or ever had any choice about whether p" (van Inwagen, 2002). Some draw in other general rules, e.g., agglomeration (a property holds of a conjunction when it holds of both conjuncts), to work α, β, □, and N over at a more complex level.[21] But the greater the complexity, the more difficult to follow error's scent.

Master-CON adds α, β, □, and N, but uses the other terms from CON, L, P_0, and P. Basically, Master-CON concludes that J's choice never was a *choice*:

1.	□((P₀&L)⊃P)	determinism
2.	□(P₀⊃(L⊃P))	1, propositional logic (exportation)
3.	N(P₀⊃(L⊃P))	2, α
4.	NP₀	Past-Fixity premise
5.	N(L⊃P)	3, 4, β
6.	NL	Laws-Fixity premise
7.	NP	5, 6, β[22]

Premises 4 and 6 involve implicit *a priori* appeals to Fixity. Old objections that hinge on

---

[21] See Kapitan (2002) for an extensive review of all such versions of master CON.

[22] This adaptation is from Kapitan (2002), pp. 148-49.

CON's "renders" and J's performative nonlocutionary omission don't apply, but new objections add to the over-determination of CON's invalidity. Those old objections that do apply to Master-CON, I assert here in the abstract. Three new objections apply.

Premise 1 only fortifies my Uniformity-based objection about CON's fixed/variable equivocation, for 1 is equivalent to its contrapositive 1',

1'. $\Box(\sim P \supset \sim(P_0 \& L))$.

As 1' makes clear, if ~P is assumed, then *necessarily* $\sim(P_0 \& L)$ must be assumed.

Rule α is "$\Box p \therefore Np$". Though α is seen as innocuous and most objections go to β, α is invalid: $\Box p$ does not entail Np. To think it does is to commit a modal fallacy. Counterexamples to α are cases of necessity involving baptisms such as "*This* is the standard meter" or performative utterances such as "I *do*" (controlling bachelorhood), or self-indexing utterances, "$\Box$(I exist)" or "$\Box$(In uttering 'I exist', I exist)". Here "$\Box p$" rests on context of utterance, e.g., what speakers self-reflexively utter or baptize, where "Np" is false, so α fails.

The reply that these are hard-determined, so nobody controls them, begs the question. To defeat these cases *non-circularly*, hard determinism must be held in the wings; an independent ground must be used, but there are no independent reasons to reject the widely-held view that certain uttered reflexives, indexicals and performatives are necessarily true, and their truth conditions are in the hands, mouths, or minds of their utterers. Since these are necessarily true ($\Box p$), yet control over their truth is in their utterers' hands (~Np), α fails.

These are special cases, as are indexicals that constitute counterexamples to the general rule that *willing or believing p cannot make p true*, but they show α is not sacrosanct. We have seen how easy it is even for the *axioms* in this debate to be misconstrued, so that is enough to warrant assault on α. But there are other grounds on which to oppose α. For

reasoning about the *actual* world, state descriptions, and alternate possible worlds involves indexicals, and indexicals pose subtle problems in interpretation regarding possible-world semantics, and more so in interpreting intensionality-containing propositions (e.g., propositional attitudes) across possible worlds.[23] We have seen errors in reasoning *across* possible worlds that was supposed to be contained *within* each possible world. Actions have *quasi-intensional* properties by virtue of their intended or teleological components (Davidson, 1968); thus, like all other *performatives*, they may also be things that we can control. It is no accident that self-indexing is linked to autonomy.

A third objection is about the detachment of the modal conclusion from Master-CON, but I will only sketch the main idea. There is a sense in which the modally-embedded proposition in the new conclusion is intuitively contingent, so suspicion immediately arises in connection with its modal prefix. The basic idea is that to the extent that the argument concludes agents have no control over their actions, it wrongly cuts or detaches the modal property "N" from the context of its relevance, which is the premises; i.e., non-control or unavoidability is not fully transitive. In the billiard ball version, greater abstraction shows the illogic vividly. With the terms from BB CON, but also Master-CON's "α", "β", "□", and "N", we have BB' CON:

BB1'.	$\Box((P_0*\&L)\supset P*)$	determinism
BB2'.	$\Box(P_0*\supset(L\supset P*))$	BB1, propositional logic (exportation)
BB3'.	$N(P_0*\supset(L\supset P*))$	BB2, α
BB4',	NP_0*	Past-Fixity premise
BB5'.	$N(L\supset P*)$	BB3, BB4, β
BB6'.	NL	Laws-Fixity premise
BB7'.	NP*	BB5, BB6, β

The conclusion is that the 8-ball doesn't go into the corner pocket and "nobody has ever had

---

[23] See Kitcher (1987) on the indexicality of "actual" (i.e., *this* world).

any choice about [it]", which is false, for the pool-player's pool-stick did.

Past-Fixity of P_0* only entails NP* (β) under the assumption that no control (N) was present, but in proximal cases it is, so the move is suspect. Control *may* be transitive; non-control need *not* be: A husband may not control his wife, and she may not control her daughter, but he may still control the daughter, and whether he does depends on their relationship, not on modal transitivity or determinism. Thus, β fails. I didn't play earlier in the game, so I had no control over the first few moves my partner and our opponents made, but when it is my turn, I do control where the balls go, depending on where I strike them, how hard, etc. As with agents, 8-balls have legitimate causal powers. When they strike other balls, they move them in lawful ways. Striking a ball a certain way *controls how it moves, despite everything in its past being out of the range of control.* An 8-ball never needs to go into the corner pocket for it to be true that *if it were struck a certain way, it would.*

Indeed, Newtonian physics – *the* paradigm of determinism – uses examples such as these as its *paradigm* cases. It is no different with J. If the causes feeding into J's decision yield J's not raising his hand, then different causes acting on him could have led him to raise his hand, equivalent to the billiard ball's being struck differently and so moving differently. Determinism requires that people/8-balls have counterfactual features. *The falsity of actualism is more an intuitive consequence of determinism than hard determinism is, rather than something hard determinists can use to reject soft determinism.* N-operators and the like only reveal CON's fallacious components more vividly. *Contrary* to actualism, a grain of salt is soluble even if never placed in liquid. The claim that it *can* dissolve, contrary to the fact that it is never placed in liquid, does not entail that it has Fixity-violating ability.

Rather, L *guarantees* that it dissolves if placed in liquid. Counterfactuals about

solubility provide the content for L and L provides the content for determinism, so hard *determinism* cannot reject counterfactuals. Conversely, since actualism denies nomic generalizations – laws – about solubility, it must equally deny the more global agglomerative generalization – L – summing all laws required to cover all state descriptions, which latter is *the core content* of determinism. The snake eats its tail. Thus, actualism is a *reductio ad absurdem* entailed by CON: *If actualism is true, determinism is false.*

Scientific reasoning and common sense concur: Salt need not be *placed* in water to be *soluble*, billiard balls need not *go* into corner pockets to be *able* to, and agents need not *lift* their hands to be *able* to. If one could *consistently* embrace it, actualism is still in need of more support than *hard or soft* determinism, and so may be rejected on that ground alone.

What justifies "N" with P_0&L is the idea that P_0&L be held fixed. But deterministic reasoning has no such implication. Whatever sense in which P_0&L is necessary, if any,[24] is irrelevant or wrongly blends necessities, in which case Master-CON cannot get off the ground by way of α, and so neither by β. It doesn't matter whether α and β are valid though neither is, despite much debate over β and agglomeration, since neither are invited unless "□" is, but "□" isn't. The whole "technical turn" in Master-CON just pushes intuition out of reach. Np is just false. Countless true counterfactuals identify agents' control over their actions without requiring *agent-access* to alternate worlds, as PAPW shows.

A more fine-grained analysis suggests CON shifts between attributive/referential senses of its terms, to borrow Donnellan's (1975) distinction. If one points at Jones, the suspect at Smith's murder trial, and utters "Smith's murderer is depraved", but Jones did not murder Smith, then one uses a definite description to pick out Jones which is false of him.

---

[24] The idea that the necessity in CON isn't transitive, in Slote (1982), defeats CON, in my view, though objections to the technical part of Slote's argument have postponed its funeral.

This use of the term is *referential*, to refer *to Jones*. One may use the same term to attribute to *whomever it is* who could have done such a thing the property of being depraved. In this usage, the definite description is used to *attribute* depravity to *whomever* it is that satisfies the description. When used referentially, "Smith's murderer" rigidly designates Jones in all possible worlds in which he exists; the attributive usage is non-rigid, and applies to whomever satisfies the description.

Under the coarse-grained analysis, P is *varied* while P_0&L is held *fixed*. Under the fine-grained analysis, "actual", an indexical, plays a role in determining the world for which "P" and "P_0" are the state descriptions. To suppose ~P is to refer to a *non-actual* world in which ~(P_0&L) is true *unless one confuses attributive/referential uses* for "P" and "P_0&L" and loses track of the properly indexed world. "What happened at T_0" can referentially designate P_0, or attributively designate P_0 in the P-world or ~P_0 in the ~P-world. So, the supposition ~P is in the *non-actual* ~P-world, and in *that world*, the supposition ~(P_0&L) is true by Uniformity. "P" first *referentially* picks *that state description determined by J's omission at T*, and "P_0&L" first *attributively* picks *whatever* state description and laws are true at T_0.[25] But once this *attributive* use is employed, P_0&L is *thereafter* taken to refer *rigidly* to that unique conjunctive proposition, heeding Past and Laws Fixities, while P is treated variably as ~P, ignoring Present-Fixity.

Using both terms attributively, "what J does at T" picks out *whatever* J does at T; "P_0&L" picks out *whatever* was true at T_0, both *in the same world*, in which case no matter

---

[25] Van Inwagen says: "It should be emphasized that '*P*' does not *mean* 'the proposition that expresses the state of the world at *T*'. Rather, '*P*' *denotes* the proposition that expresses the state of the world at *T*. In Kripke's terminology, '*P*' is being used as a *rigid designator*, while 'the proposition that expresses the state of the world at *T*' is perforce non-rigid." (1975, p. 53; footnote omitted) See Kripke (1972). In my terminology, the former is referential, and the latter is attributive. Oddly, van Inwagen makes this distinction, yet fails to see his equivocation.

what world we entertain at T, T_0 is in *that same world*, by Uniformity. *Using both terms referentially,* P picks out J's actual omission, and P_0&L picks out the unique past/laws that obtained at T_0 in *that same (actual) world.* Our *unproblematic symmetrical usage* here *proves* equivocation.

CON implicitly assumes Fixity for P_0, and its *apriority* is persuasive, but masks an error. Adler (1995), for example, insists P_0 is unfalsifiable because it is an arbitrarily selected prenatal (non-rigidly identified) state description with no content, as if it *means* '*whatever* state description is true at T_0' – ignoring van Inwagen's remarks about its rigidity. On his view, P_0&L would be *analytically* true, because no matter what else is varied, P_0&L picks out *whatever* is true at T_0. I call this faulty blending of necessities, contrary to Slote's caveat, formalized in α, the *"fallacy of homogenizing necessities"*. α seems innocuous because it involves modal *weakening* from $\Box p$ to Np, analogous to weakening from $\Box p$ to p, which seems indisputable; but it is noxious.

Holding "P_0&L" tautological/unfalsifiable commits the fallacy of persuasive definition. "P_0&L" is then held synonymous with "*whatever form* Fixity requires". But if we also held "P" analytic so it remains true though J *does* raise his hand, contracausalism is unfalsifiable, and *freedom is analytic too, for – necessarily – doing otherwise violates no Fixity rules!* When *only* P_0&L is seen as *fixed*, but J's ability to vote for or against is seen *variably* as mapping onto P or ~P, P is seen in the end as *fixed*, as if proof that ~P and freedom are seen together as impossible.

It is not sensible to attribute depravity to *whomever murdered Smith* while supposing Smith lives. The attributively-identified murderer cannot exist *as such* in the same world in which Smith lives, since Smith's murderer's existence *as such* entails Smith' murder.

Likewise, we cannot attribute to P_0&L that it is *whatever entails P* in a world in which ~*P* is supposed. For while Smith's murderer *can* exist, though not *as such*, in a world in which Smith is not murdered, P_0&L *cannot* exist or hold in the ~P-world. Because "P" and "P_0&L" are so indefinite and their world-selecting functions are so subtle, the equivocation is opaque.

Since "P", "~P", "P_0&L", and "~(P_0&L)" have been treated ambiguously, lets's tag *actual* and *counterfactual* alternatives for each with distinct symbols:

> Let $R_1,R_2,...R_n$ denote *actual-world* state descriptions; and
> Let $S_1,S_2,...S_n$ denote *counterfactual-world* state descriptions (say, in the nearest possible world), all in parallel with $T_1,T_2,...T_n$, which model temporal succession.
> Let L_1 denote the *actual-world* set of all natural laws previously denoted by "L";[26]
> Let $L_2,L_3,...L_n$ denote *counterfactual-world* sets of natural laws; and
> Let the order of $L_2,L_3,...L_n$ denote *increasing distance* from L.
> Let W_1 denote the *actual world*;
> Let $W_2,W_3,...W_n$ denote *counterfactually possible worlds*; and
> Let the order of $W_2,W_3,...W_n$ denote *increasing distance* from W_1.

We stipulate that "P" *attributively* picks out R_i at T, and "P_0&L" *also attributively* picks out R_1&L_1 at T_0; both are true in W_1. Now, *suppose* J raises his hand at T. If so, J is in W_2. Observing Uniformity, such that *both* "P" and "P_0&L" *attributively* pick out *whatever* state descriptions are true at T and T_0, respectively, in W_2, we stipulate that the state description at T in W_2 is S_i, which entails ~R_i in W_1, and at T_0 it is S_1&L_1,[27] which entails ~(R_1&L_1) in W_1.

Now, if P and P_0&L in W_1 denote R_i and R_1&L_1, respectively, and R_1&L_1 entails R_i, then to suppose ~P is to suppose ~R_i, which by Uniformity takes us into W_2, where J's act is represented by S_i, for which the prenatal antecedent is S_1&L_1. If we read *both* P and P_0&L attributively as we switch from W_1 to W_2, there is no problem, since the claim *that S_1&L_1 entails S_i* is Uniformity-equivalent *in W_2* to the claim *that P_0&L entails P*, i.e., that R_1&L_1

---

[26] "L" becomes "L_1" so the subscripts will all match for ease of bookkeeping.

[27] This is so if the pragmatic aversion to varying the laws is cogent. But if not, the same apparatus could be modified by altering the nomic component but holding fixed the historical component, e.g., R_1&L_2, or by altering both. Only pragmatics hinges on this.

entails R_i, *in* W_1. There is no problem with the symmetrical referential reading. The problem arises when we shift from one reading to another *across* worlds – illicit by Uniformity.

CON shifts from one reading to the other, as if *two worlds'* state descriptions were in *the same world.* First, "P" *referentially* picks state description R_i (about J's omission) at T in W_1. Second, P_0&L *attributively* picks R_1&L_1, the state descriptions for *whatever* is true in W_1 at T_0, and then R_1&L_1 is *referentially* rigidified by Fixity. Third, supposing J can do otherwise, we shift from R_i (from "P", about J's *actual* omission) at T in W_1 to S_i (to "~P", about J's *counterfactual* action) at T in W_2, *another world.* Finally, here J's doing otherwise – S_i in W_2 – is seen as impossible because it entails W_1's ~(R_1&L_1), the referent of "~(P_0&L)", seen as impossible by Fixity, rather than as an error in worlds-bookkeeping.

For here we end up with state descriptions from *different worlds* in the same domain of discourse, R_1&L_1 from W_1 at T_0 and S_i from W_2 at T, *proof* of equivocation. Since S_i entails ~R_i, but R_1&L_1 entails R_i (master-CON-sanctified as "□(R_1&L_1)⊃NR_i" (by α's modal weakening from □ to N), ~R_i (and ability to do otherwise) seems impossible. Again, CON's error is to pick R_1&L_1 as *whatever* state description at T_0 in W_1 entails (as with P) R_i at T, rigidify R_1&L_1 by Fixity, shift from W_1's R_i (J's omission) to W_2's S_i (J's action), and then rule out S_i because R_1&L_1 (from W_1) – held fixed *across* worlds by Fixity – entails the contradiction ~S_i. S_i and thus conditional ability are seen as impossible, but the contradiction rests on equivocation.

An adjustment to our notation might help clarify my claims. Let us use "S" as a variable whose values are either P or ~P; which value S takes depends on the context of utterance. If in a given context the subject of discussion is world W_1 (which is the same as our world, W), then S is P; if the world is W_2, then S is ~P. Let "P_0" be a variable whose

values are $P_{0,1}$, $P_{0,2}$,..., P_n. Depending on which world is under discussion, $P_{0,i}$ is the state of the world at time 0 in world W_i. Let "L" be a variable whose value is fixed by context; when in a context we are discussing world W_i, the value of "L" is L_i. L_i are the laws of nature in world W_i. We may refer to the laws of a world as "L" when greater specificity is not required. We may assume the same times in all worlds, to avoid complications. Thus, my claim is that "$\Box(P_0\&L)\supset P$" is always true if the context is constant – if a single world is discussed. But CON shifts the context for P by moving to W2, and when it does so it faithfully treats the consequent, P, like the variable it is, as an *attributive designator*, but when it gets to the antecedent, P_0, it keeps the context fixed at W1, thus treating "P_0&L" as if it *rigidly designated* $P_{0,1}\&L_1$. This is asymmetrical, equivocation.

This analysis plays a role in a determinist account of possibility that comports with PAPW. I will spell it out in full in chapter 5, but sketch it here. When we talk about Agent's ability to do otherwise, we do *not* mean he has *access* to W_2 or can bring about ~P in W_1, as if in choosing he is *determining*, of many worlds he has access to, *that* world which he shall by his decision *make* actual; Fixity rules *that* out. We only mean that *had he wanted otherwise* – which would put him and his past in, say, W_2 – he'd have done otherwise, since he'd have been in W_2, and there he must do otherwise, given Uniformity. His actions are *just as necessitated* by his *different* past in W_2 as they are in W_1. PAPW2 implies only that *to discern* whether it is true that *had he wanted otherwise (i.e., in W_2), he'd have done otherwise (in W_2), we* – not Agent – need to *refer* to W_2, and maybe W_3 and W_4, where we *suppose* altering the variables: opportunity and motive. We are *supposing* the use of Mill's methods across worlds. To do that, *we refer* to the nearest possible world and vary, say, only his *motive*, and 'see' – *infer* from counterfactual reasoning – whether in that world he does otherwise; if so, he was free to not do what he did, had he wanted not to do it (PAPW2);

alternately, we vary only his *opportunity* (PAPW1). *Agent accesses no alternate worlds.*[28]

We may say that PAPW puts at least three possible worlds in play: W_1, the actual world in which J does what makes it the case that P (i.e., J doesn't want to and so doesn't raise his hand); W_2, in which an opportunity that was blocked in W_1, so that J can raise his hand if he wants to, given different antecedent conditions, i.e., $P_{0,2}\&L_2$, but in which he still does not want to, hence does not raise his hand; and W_3, if the removal of the previously-blocked opportunity and the new antecedent conditions ($P_{0,3}\&L_3$) make it such that J can raise his hand and J wants to raise it, and J does raise it, then ~P is the value of S in W_3, satisfying PAPW2. Three worlds are minimally necessary to specify PAPW-satisfaction, but more may be needed to identify more complex motive/opportunity pairs and failures.

In both W_2 and W_3, the new antecedent conditions include the removal of the previously-blocked opportunity, but in W_2, J still doesn't want to raise his hand, whereas in W_3, J does want to raise his hand; in this way, we isolate external bars to the expression of J's intentions, and test J's behavior in hypotheticals with and without those *external* bars in place. This is a hypothetical version of a Millian causal analysis, and it may be developed to isolate J's *internal* bars to the expression of his intentions as well as his internal ability to refrain from their expression.

J need never 1st-personally access alternate worlds, not even in thought, though he *may entertain them* in Bok-style deliberation. Likewise, a knower need not know he satisfies Nozickian truth-tracking conditions to satisfy them, though he *may*. Rather, we simply 3rd-personally consider, from an externalist, counterfactual perspective, whether J's intentions are effective under the counterfactuals described so we may infer whether he is/isn't

---

[28] This is just a thought-experiment, analogous to the one Nozick (1981) imagines performing on cognizers to see if their beliefs are truth-tracking. I defend this view in chapter 5.

externally/internally blocked in the expression/effectiveness of his intentions.

So, to identify the causal role of a blocked opportunity in W, we need to look at worlds W_i and W_j in which the previously-blocked opportunity in W is now open. J doesn't want to raise his hand in W_i, and he does want to in W_j. We use these worlds to determine whether his desires are effective under world-conditions in which their expression is not blocked externally. To test the role of opportunity, we need at least three worlds, where actual world W is included in the count. To number these worlds, in W_2, an opportunity is open that was closed in W_1; if he still does not want to, and doesn't do so, the fact that the newly-opened opportunity did not result in his raising his hand evidences that the blocked opportunity in W_1 was not the cause of his non-raising of his hand in W_1.

I wish to avoid giving any impression of the very thing I do not want – accessibility to the impossible. I do not want to lapse into saying "if the world had been different, he could have done X" without the same analysis for "could have done" in the *new possible world* as the analysis for "could have done" in the actual one. So, we must imagine the initial conditions and laws changed to $P_{0,3}$. So, by "in W_2 he could have done what brings about ~P", I mean that, if the state of the world as it was at an earlier time is varied in W_2 so that we are in a world W_3, where $P_{0,3}\&L_3$ hold, then in that world J does raise his hand, and ~P holds. That is consistent with saying in W_3, $P_{0,3}\&L_3$ hold as well, by Uniformity. This is just what is needed to defeat CON.

PAPW1 *refers* to W_2 to 'test' the nearest world where Agent is *able to do otherwise*, i.e., where an opportunity is open there that is closed here, but he has the same motives in W_2 as in W_1; PAPW1 refers to W_2 to see whether he does the same thing *when given the*

opportunity.[29] If he does the same thing in W_2 *though able not to*, then his act in W_1 is voluntary. But if in W_2 he *does do otherwise*, then we may infer that his lacking that option in W_1 was causally responsible for his action in W_1, so he did not act fully voluntarily in W_1.

Technically, "able to do otherwise" in W_2 requires us to refer to W_3, where not only is the opportunity present, but also his altered desire. PAPW requires that *we consider* – not that *agents* have *access* to – at least a triplet of worlds:

W_1: the actual world, where the alternative possibility is blocked (by Frankfurt-style counterfactual interveners or simply by determinism), and he does X;
W_2: the nearest possible world in which the alternative is unblocked (by absence of a counterfactual intervener or by different history and/or laws), but not used (he does X, and he doesn't want otherwise, though he is able to do otherwise, should he want to); and
W_3: the next possible world (represented by the unblocked alternative in W2) in which the unblocked alternative is taken (he wants to do otherwise and does otherwise).

This *externalist* analysis identifies agential causal efficacy.[30] Master-CON suggests non-control due to *prenatally-transitive inability*, but is also fallacious because PAPW shows how agents have control *without* agent-access.

Little remains to say about CON. Van Inwagen challenges the conditionalist to replace statements of form (8) with those of form (9), claiming that either no sentence in CON would turn out to be false or, if it did, that would itself disprove the translation (1975,

---

[29] The nearest world in which these conditions hold may not be the nearest possible world, but we call the nearest world in which they do "W_2". In the application of these principles to real agents, the distance between the actual and nearest possible world in which the conditions hold matters, as we'll see in chapter 5.

[30] Fischer (1987, in Pereboom, 1997a, p. 236) and F&R (1998, p. 53) defend a similar view with "actual sequence" *reason-responsiveness*. On their view, Agent need not access an alternate world if his deliberative-mechanism is reason-responsive. We *imagine* varying the circumstances; e.g., *were* a stronger reason present that should defeat Agent's choice on the assumption that Agent is rational, either Agent's mechanism would/wouldn't respond to it. If yes, then Agent's mechanism is reason-responsive; if no, it is not reason-responsive. Testing for reason-responsiveness doesn't require that *Agent access* alternate worlds, just *the judgment* whether the mechanism *would* respond differently under better/worse reasons. F&R hold reason-responsiveness necessary and sufficient for responsibility without requiring alternate possibilities. Hence, theirs is "semi-compatibilism": Determinism is compatible with moral responsibility, but not with alternate possibilities.

p. 57). But the (9)-like version of (7),

(7') If determinism is true, if J had chosen to raise his hand at T, J could not have raised his hand at T.

is false, for choices are not impotent, but potent, under determinism – choices, e.g., A, are *what determine* state descriptions, e.g., P. So, van Inwagen's "let the reader do it" bluff fails.

One final challenge to our view comes from Lehrer (1966), who argues that conditionals cannot capture the meaning of "can" because the following, which I call "Trio", is consistent:

1. If C, then S X's.
2. ~C.
3. If ~C, then S cannot X.

Lehrer's argument is not taken to target ordinary cases of ~C, where S's not wanting to raise his arm is what keeps S from doing so, but unusual cases in which, say, the motor nerve from S's brain to his arm is blocked, but would open upon S's wanting to X; normally, the nerve is open regardless. This unusual case is "finkish" (Martin, 1994), and applies to dispositions such as solubility: Imagine a spoon of insoluble 'salt', and an angel standing by ready to change its structure if it is immersed, enabling it to dissolve.

This 'salt' lacks inherent solubility, though the conditional, "If this salt is immersed in water, it will dissolve" is true, thanks to the angel. This divorces the antecedent from the conditional for the disposition. Satisfaction of the antecedent is taken to obliterate the disposition: "If this salt is immersed in water, it will dissolve" is true, but "this salt is soluble" is false. But 'salt' is not salt if not soluble, but is functioning *as salt-like* only in a purely philosophical-fiction – "phi-fi" – sense less innocent than that of Twin-Earth-scenario counterparts. There, a functionally-equivalent chemical structure for "tsalt" – analogous to Twin-Earth 'twater' – is soluble. But our phi-fi salt – "phsalt" – plays a mischievous role.

Tsalt/salt is naturally soluble, but phsalt is naturally insoluble, so cannot function as a dialectical foil for intuitions about dispositions of naturally-soluble substances like t/salt. To treat phsalt as if it functions this way presupposes that the phsalt/salt relationship is like the tsalt/salt or twater/water ones, where internalist or "narrow" epistemic strategies cannot distinguish cognitively-identical cases only distinguishable externally or "widely". But these pairs are disanalogous: The phsalt/salt pair is not cognitively-identical, for both tsalt and salt are naturally soluble, but phsalt is supernaturally soluble. It doesn't matter that phsalt is not soluble except under an additional angelic condition, for other substances' dispositions also have more than one condition of exercise. With phsalt, we need to include the exercise condition, "and an intervening angel is standing by determined to transform it on contact with water", in C. We can say that t/salt is naturally soluble, and phsalt supernaturally soluble. The problem is the result of treating phsalt's and t/salt's solubility conditions inconsistently in interpreting Trio – an equivocation.

Phsalt cases are not tsalt/salt cases, and phsalt is phi-fi-soluble or "phsoluble" anyway, so long as the angel's disposition to render phsalt soluble is part of C and is as reliable as salt's micro-structural configuration is in its disposition to dissolve on contact with water. The supernatural and natural conditions must be applied consistently.

A distinction between "ability" and "capacity" may be described as ability in *narrow* or *wide* senses, to avoid ordinary-language connotations. Usually, both abilities coincide, but in phi-fi cases, they may diverge. I argue that *when* – note *temporal* indexing – not immersed in water, phsalt lacks *narrow* ability to dissolve, but possesses *wide* phability, since *if* dunked in water, it acquires narrow ability, and even when not dunked *can* acquire the narrow ability implicit in phi-fi-fink cases. Salt lacks neither narrow nor wide ability. If we are consistent about which sense of ability is being used, we can conclude that Lehrer

equivocates between the different senses in interpreting his premises. He also fails to index time consistently in terms of when C is/isn't satisfied and/or when C includes phabilities or just abilities. Consider "example A":

A1. If a student gets a 90 or above, then he gets an A.
A2. The student does not get a 90 or above.
A3. If the student does not get a 90 or above, then he cannot get an A.

As Lehrer's challenge is generally interpreted, he is imagining a case in which failure to achieve a 90, let's say, finalizes a precarious phi-fi state of the student's nervous system that leaves him incompetent to understand philosophy. But the student still retains the phability to get an A. Lehrer's puzzle applies only to phi-fi cases, but fink cases are *non-representative* of what goes on when we distinguish between abilities in broad and narrow senses, which is enough if not under the spell of actualism. Conditionalists need only note that Lehrer equivocates on what the relevant ability or disposition is: If the disposition is ordinary, then the consequent of A3 is false; if it is extraordinary, like the power to dissolve under the auspices of an angel, A1 is false if taken to express non-phi-fi ability.

Ph/salt *can* dissolve *when* in liquid, *when* C – with/without angelic content – is satisfied; and it *cannot* dissolve *when* not in liquid. Salt cannot dissolve midair. It requires contact with a solvent, but it's still widely soluble; aluminum is not. The phi-fi take on this is that *when phsalt is in midair* it is insoluble, for phsalt-extrinsic angelic agency alone can transform it. I've agreed that phsalt is *narrowly* insoluble *when in midair*, and acquires ordinary narrow ability *when in water*. But even in midair it has the extraordinary narrow ability to "phissolve" – to dissolve if put in water *and* touched by an angel – and the extraordinary broad ability to acquire the narrow ability; it will acquire the capacity to dissolve, and dissolve, if put in water.

This applies to my capacity to raise my arm. I am narrowly able to raise my arm

when I want to; a paralyzed person is not when he wants to. Salt can dissolve on contact with water; aluminum cannot. A light bulb can go on, but only if the switch is on and current flows into it. It cannot go on *per se*, unconditionally, when the switch is off, or the wires are damaged; a dead bulb cannot go on under any conditions. Some abilities are phabilities; others are finkish in natural ways, e.g., grades involving plagiarism; others are ordinary. So?

This can be put in either of two ways: Either (1) given that the angel has interceded, it has narrow ability, or (2) even before the angel intercedes, it has broad capacity because, *if the angel intercedes*, then it would dissolve. (1) is consistent with actualism; (2) implies non-actualist counterfactualism. I may not have the independent power to teleport myself, but should a sci-fi hand-held teleportation device transmit that power to me, I may. Supposing the hypothetical, I have the power counterfactually. A rock cannot dissolve in water, but should a 'phalchemist' transform its structure just as it makes contact, it will; supposing the hypothetical, it has the counterfactual power. Some objects might lack even the counterfactual powers, say, if their elements are not amenable to phalchemy or if God made a counterfactually-insoluble rock. If Lehrer objects that we are supposing too much, my reply is that we are entitled to the same dialectical privileges. Supposability is tricky, but not only for us.

Capacities-for-capacities complicate matters, but not much: Infants lack the capacity for continence *as infants*, but not the capacity *for* the capacity later on. *The way* continence *as a developed capacity* is determined is *heterologous* with respect to *the way* its exercise is determined. To blend both due to *mere deterministic necessity*, as if both equally imply N (by α), is to commit the homogenizing necessities fallacy. The more complex the ability, the more relevant to freedom. Some capacities involve restraining motivational subsystems, say,

primal desires. Consider "example C".

C1. If S wants to restrain his otherwise runaway desires, then S so restrains them.
C2. S does not want to restrain his runaway desires.
C3. If C2, then S cannot so restrain his runaway desires.

Though C3 is analogous to being unable to raise one's hand when one doesn't want to, here is a complexity absent in the arm-raising case. I may want very strongly *not* to refrain from acting on my runaway gustatory or sexual desires, because refraining will prevent great pleasure, yet my Kantian commitments to my diet-buddy, to whom I have sworn an oath, or my life-partner, may be sufficient for me to refrain *despite the fact that I don't want to refrain*. Here, my not satisfying C1's antecedent is insufficient to defeat even *my narrow ability*. C1 and C2 are consistent with my restraining my runaway desires. And that is inconsistent with C3, in which case Trio is in error to place a modal in 3, for there are false instantiations of 3, e.g., C3.

I speak English effortlessly, but Italian only with effort; bilinguals speak two languages effortlessly. Some capacities require more exercise conditions than others, and some are linked; the Kant example above links with the capacity to act on duty. The more complex the capacity, the less sense 3 makes. But 3 even makes no non-phi-fi sense with simple solubility, so long as 1 is held true only when its antecedent is *causally sufficient* for its consequent. Freedom involves many capacities like those illustrated in example C. There it is clear that agents may not want to restrain their desires, but that is *not* a good reason to think they *cannot* do so. Lehrer might agree, but insist on the rare case in which not wanting to restrain desires *is* sufficient for being unable to restrain them, for instance, because a desire to restrain one's desires is necessary for a finkish state of the nervous system that gives one the capacity to refrain. That's the counterexample.

But this case is not phi-fi, for something like it occurs in a penumbral range of akratic

cases where agents feel powerfully seduced by the immediate prospects of the gratification, and *experience inability* in the absence of countervailing occurrent reasons: Were countervailing occurrent reasons in favor of restraint vivid, they might tip the neuromuscular scales and trigger the ability, but in their absence, Agent *experiences* the proprioceptive cognitive/conative phenomenology of inability. Here, the inference of *narrow* inability is very intuitive, but there is no threat to the capacity/ability distinction here, nor to the idea that the narrow ability would appear if its conditions of exercise – normal, phi-fi finkish, or just plain finkish (as here) – were present. Let us simplify this interesting but unproblematic case, a 'plain finkish' instance of Trio:

PF1. If I want to refrain, I will.
PF2. I don't want to refrain.
PF3. I can't refrain.

PF1 shows that I have the ability in the broad sense; if I want to refrain, that sends my nervous system into the state that normally subserves the ability to refrain. But since I don't want to refrain, my nervous system remains in an akratic state that does not subserve the ability. This requires no reference to my other ideas. Lehrer equivocates on which ability is being referred to.

I have taken pains to identify the homogeneity fallacy in α, the erroneous transitivity of non-control in β, actualism, CON's needlessly strong view of possibilities, and CON's equivocations. These flaws are implicated in both theological and logical fatalisms. Since my primary concern is easy-autonomy, I only sketch the connections between hard determinism and its fatalistic cousins here.

It's easier to dismiss fatalism, since it is more tautologous than hard determinism. As for God, our choices determine God's beliefs about them, as J's choice determines P.

Supposing we *control* our choices, this determines God's beliefs, not vice versa: "Theodoxastic-Fixity" doesn't undermine choice, but is a function of omniscient truth-*tracking*, not truth-*determining*. That is a problem for hard, not soft, theological precognitivists. The logical fatalist, e.g., Bernstein (2002), likewise, cannot square the timeless view of truths, "Alethic-Fixity", with the fact that those truths contain information about how agents control their actions. If agents control actions, this determines what the timeless set of truths will say about them, not vice versa, just as the fact that J's psychoneural states determine the content of P at T, not vice versa. Both fatalisms are 2-card montes, unlike hard determinism's 3-card ones; determinists' fatalistic leanings at least rest on *substantive* confusions about the doctrine's *causal-nomological* content.

Hard determinists reason *backward* to Past and then back to Present-Fixity, but *logical* fatalists reason *outward* to tenseless truths and thus to Alethic-Fixity (by α), and then back to Present- and Future-Fixity. This is analogous to hard determinists' ostrich psychology, viewing local facts from the global vantage, from which angle causal control is not visible. *Theological* fatalists, e.g., Zagzebski (2002), either reason *backward* to God's precognitive states and thus to Theodoxastic-Fixity (by α), or *outward* to where God's timeless omniscience accesses logical fatalism's timeless truths and thus Alethic-Fixity (also by α). But Alethic-Fixity's entailment of Present- and Future-Fixity, and their entailment of N (by β), is what allegedly robs us of freedom, yet we have shown α and β invalid. Present- and Future-Fixity are consistent with autonomy, if actualism is false.

The fact that both Theodoxastic/Alethic fatalisms share these features with hard determinism is a dialectical embarrassment, for this reveal that these errors have *nothing to do with causation*, whereas we began with the concession to hard determinists that their

grounds for confusion might be excused by reference to substantive matters having to do with causation. That ground is removed. All three pessimisms are multiply-fallacious.

CONs formed with P_0 for God's belief or timeless truth, supported by Theodoxastic- or Alethic-Fixity, equally fail. In the *original* fatalist CON, we begin with P and the supposition that ~P representing J's ability to do otherwise, and then show how that violates Theodoxastic- or Alethic-Fixity of P_0, and so is impossible. In the *master* fatalist CONs, α implies that God's belief/timelessly true proposition $\Box P_0$ entails NP_0, which by β entails NP. Original and Master-CON are both invalid, and since both fatalist CONs share their essential features, they are also. Thus, agents need not *violate Present-Fixity, bring about ~P or access other worlds* to establish their control over their actions, and fatalistic actualism – theological or logical – doesn't change this. There is *less* reason to think God or logic constrain actions than to think causation does, so both fatalistic CONs are more idle than both deterministic CONs. But that CON's illogic applies to both fatalisms proves CON's errors are not deterministic, *per se*.

This chapter has sought to dispel horizon-level pessimism. I conclude that all forms of pessimism are either false or implausible, and libertarianism is unmotivated. This leaves only *soft* in/determinism. Like Mele (1995), I think autonomism is more credible than either determinist or indeterminist non-autonomism, though I need argue only for determinism for easy-autonomy; absent strong empirical evidence, I need not appeal to indeterminism. Since hard determinism poses the greatest threat to autonomism, to dispel that threat on as minimalistic grounds as are necessary, it suffices to show that no questionable assumptions are needed to dispel pessimism.

The arguments in the remaining chapters, then, are limited in scope, mostly *internal*

to soft determinism, to the easy problem of autonomy. Let us turn to them now.

Chapter Three: The Off-Line Metamental Architecture of Autonomy

Chapter 3 identifies a certain feature of metamental states constitutive of easy-autonomy, namely, their relatively-*off-line* character, and models their causal/functional properties in illustrative cases drawn from a variety of sources. Certain primate-simulation skills are *one possible* evolutionary source of these abilities; others include mammalian dreaming and language-use. In a *contrastive* sense to be explained, the organism engaged in these activities operates its usual I/O, S/R, cognitive/conative mechanisms 'off-line' – and thus *partly-endogenously – relative to* their *typically-unmediated* on-line engagement with the environmental causal stream, which latter involves *directly-exogenous determination.* It is only *relatively* off-line, and is still *ultimately* on-line, for everything physical is on-line relative to global determinism. I argue that relative off-line status is sufficient for autonomy.

The appearance of any ability to go off-line/metamental counts as an evolutionary *watermark* for the beginnings of "organismic self-regulation" (Perls, 1947), the maximum of which is autonomy. *Any off-line/metamental functions* involve elements on the spectrum of abilities we associate with autonomy, i.e., the ability to act on or refrain from acting on one's own desires, or to reflect on and alter one's own mental states and dispositions. Organisms that can intentionally alter their I/O states possess causal self-control powers lacking in organisms that cannot, but who, instead, are limited to immediate S/R behaviors.

The behavior of organisms limited to unmediated S/R, organism/environment interactions is fully causally controlled by exogenous forces, whereas that of self-adjusting organisms stems from mediated I/O processes that loop through the conscious, off-line circuitry of the organism, and is thus more endogenous. Compare woman W, who knows she has difficulty regulating her emotional states and takes serotonin re-uptake-regulating pills to regulate them, with woman W', who has the same problem but lacks W's knowledge

of the pill. W exhibits self-regulating powers W' lacks. An organism that can condition its own S/R pairs – through biofeedback, auto-hypnosis, or by hiring a behavior modification therapist (Mele, 1995) – has *some* endogenous self-regulative-control and is at an advantage over one *entirely* exogenously S/R-determined.

The difference between organisms with/without endogenous powers is as concrete as that between locomotory/non-locomotory organisms. Let us call the behavioral model whereby future expectations lure organisms the "pull", "teleological", or "forward-looking" model. If sentient beings never controlled their movements, the pull-model would never apply to them. Instead, a "push", "hydraulic" or "backward-looking" model would, such that their muscular contractions were pushed by propulsive forces, as is a person shoved by a crowd. To avoid stronger connotations, I *stipulatively* use "hydraulic" *only in the sense adumbrated*, only similar to *propulsive liquid-flow through pressurized systems*. If a hydraulic model were correct here, most sentient beings would suffer from universal motor paralysis, contradicted by the most basic data/theories in biology. Yet, hard determinism suggests this hydraulic model.[31]

We are aware of much of what moves us, W's awareness of this brings its possible control. As Chalmers' "bridge law" states, consciousness brings with it the power (1) to verbalize its contents, and (2) to act on its contents (1996). Consciousness brings with it other powers implicated in our self-regulative ability. Consciousness of the causation in one's system brings the possibility of its control, and its control is a functionally-distinct kind of causation that enables us, as Locke noted (1690), to stand back, disconnect our I/O, and self-regulate. Nonhuman animals lack this endogenous causal ability.

---

[31] See Stelmach (1976), McGuigan (1978), and Langer (1967, chapters 11 and 12) on the biological bases of pull-model capabilities of locomotory organisms.

We reject with CON the idea that endogenous actions being *ultimately* exogenously caused is sufficient ground to infer that endogenous organisms are not self-controlling. *Autonomy just is self-control*, so caused behavior is free if it is self-controlled. Natural selection favors organismic self-regulation; thus, evolution favors autonomous organisms. Dennett (1984) gives a "just-so" story of evolutionary developments that *may have* made reflexive powers, thus autonomy, possible. My just-so story *suggests* their possible development from primate-simulation mechanisms, for which there is much evidence.

Since the evolutionary component of my theory suggests its *off-line* element *may have arisen* from primate-simulation skills, I call it "the off-line theory". Since the off-line element is a *well-defined feature* of simulation theory, an account of simulation theory will provide apparatus to explain 'off-line'. There is a debate in folk psychology between the "theory-theory" and simulation theory, but I sidestep this with a hybrid theory. The simulation mechanisms we possess even on the weakest hybrid theory endorsed by Stich and Nichols ("S&N") (1993) involve ability to take cognitive/conative systems off-line. This ability figures in the self-regulating powers folk psychology takes us to possess.

Call the claim that the viability of the notion of an off-line function is independent of simulation *theory* the "independence thesis". Though I *suggest* primatological just-so origins for off-line abilities, and support them empirically, they are independent of simulation and simulation theory. Simulation theory is popular in philosophy of mind, so the off-line theory simply draws on the illustrative/explanatory currency in which it trades.

For the same off-line concept in *simulation* is found in computer science, hydraulics, and other systems models. The off-/on-line distinction involves *relative* terms, for a computer may be on-line *relative* to a local-area-network, but off-line *relative* to the Internet. Similarly, a thermostat may be on-line *relative* to its thermometer and the boiler to which it

is wired, but off-line *relative* to the computer system connecting it to a larger climate-control system if it is in manual overdrive mode. Off-line relativity comes in degrees; e.g., runner R1 thinks about a race calmly, but R2's muscles tense as he thinks about it.

Given that off-line functions are *relatively* off-line, no Fixity-violations arise. In rejecting β, we opened the possibility of *caused control.* Off-line-status explains how it may be *functionally* possible; it is the key to the possibility of *self-reflexive* functioning. Thus, when an organism intentionally disengages its causal 'roots' from the environmental/causal nexus, it goes off-line for purposes of conscious self-regulation. Autonomy is organismic self-regulatory ability, evolved naturally from off-line/metamental abilities. Simulation is an ability of mind – conceived non-dualistically – to replicate non-occurrent mental states – one's own/others' – off-line relative to their usual on-line S/R-engagement, so it is a *metamental* ability. *The birth of metamental abilities made possible self-reflexive ones, which made possible self-controlling ones, and autonomy just is self-regulation.*

Monkey-see-monkey-do may have led fortuitously to autonomy. The off-line element of metamental behavior and its novel looping through consciousness makes reflexivity and thus metacausal control possible. The present analysis concerns *observable* behaviors, is *evidentially pure*, takes minimal *theoretical risk*, and uses the behaviorist's *functional concepts* of *S/R, I/O*, and other simple *causal/functional* concepts that admit of simple cause/effect *flow-chart diagraming*. Thus, the analysis suffices for easy-autonomy: It identifies brain/behavioral, causal/functional features of autonomy.

Simulation theorists account for our folk success as an evolved ability to simulate minds.[32] There is a graded scale of simulation skills, with brute mimicry at the primitive

---

[32] For an introduction to the literature on simulation theory, see *Mind and Language* 7(1992).

pole and counterfactual reasoning at the sophisticated pole,[33] with gaze-following, pretense, and supposition somewhere in the middle. In simulating, I simply need to recreate the target's state in myself; no theory or mentalistic vocabulary is necessary, for example, for mimicry. We may understand, predict, and explain other minds by taking our decision-making system off-line, feeding it hypothetical inputs/intentions we lack, and seeing what our system outputs; these might serve as the basis for prediction and explanation.

Gordon (1992, pp. 87ff.) objects to equating "off-line theory" and "simulation theory" because the off-line element of simulation is ancillary. We may engage the mechanism on-line on one side and not the other (pretense), vice versa (in hypothetical and counterfactual reasoning, and in predicting our own behavior), and on-line on both sides (in our own decision-making). We may generalize from Gordon's observation. Not all the relevant behavior is off-line; nor is all off-line behavior simulational. Lets' look at some examples of simulation, to illustrate their applicability to the issue of autonomy.

Suppose I pretend that, like Fred, I desire x and believe I cannot get x unless I y, so I pretend-decide to y. My pretend-decision to y is the basis for my prediction that Fred will do y. Assuming explanation works as retrodiction, a simple adjustment explains Fred's doing y: When I simulate wanting x and believing no x without y, I do y.

When hiker B is walking behind A, why isn't B amazed that A turns, moves, goes left, up, right, etc., just as the hiking path calls for? Gordon suggests:

> You aren't puzzled ... when he raises his leg high ... when there is an obstacle to step over.... When we are aware of others – ... as others – we are constantly, automatically projecting onto them our own beliefs about the environment. (1992a, p.16)

---

[33] Walter has stated: "To express it somewhat more technically, *we mentally simulate counterfactual situations*. The prefrontal cortex generates these scenarios of future events.... Simultaneously the amygdala and hypothalamus (hormone control center) also affect one's body, particularly the visceral functions. The body reacts as it has in similar past situations and, via *feedback-loops*, it reports its state back to the brain." (2002, p. 571; emphases added)

The above is a simulational model of disengaging one's decision mechanism from one's beliefs/ desires on the input side, and from one's action-control systems on the output side.[34]

We want to know why Pierre went to Lyons. The theory-theory says that we believe that Pierre went to Lyons because he believed his lover, Adele, was there, he wanted to see her, and his belief and desire are nomologically sufficient – in his tacit theory of mind – for his going to Lyons. Simulation theory says we can know why Pierre went by simulating him: If I believe that my lover is in Lyons and I desire to see her, then I will go to Lyons. In simulating him, we take our own I/O mechanisms off-line, or else we would really go to Lyons; the output – what anyone with Pierre's input would get – is the simulated decision to go to Lyons. Since we all would get that output under that input, Pierre's behavior seems intelligible, predictable, and explainable.

In a boxing match, a warning is given for a move we did not observe. We come to grips with referee R's call by supposing R could see what we could not. To evaluate R's warning to boxer B, we imagine ourselves in R's position, the inputs that likely went into R's I/O mechanism, and see what outputs arise. We might form a belief that R's locus permits a perception of an illegal elbow strike by B not directly observable from our location.

When a motorist appears to rudely cut me off, to judge whether there is an objective basis for Strawson's 'reactive attitude' of resentment, I simulate being in her shoes, and wonder whether there are any inputs she might have, e.g., an emergency, such that, were I to have those under similar circumstances, I might behave as she does. If I would not normally feel shame for that sort of excused rudeness, I will not normally resent her driving.

In basic simulation, a 1st-personally described simulator "S" simulates a 3rd-

---

34 *Cf.* S&N (1992), figures 1 and 3, pp. 40, 53.

personally described target "T". In a variant, e.g., the rude motorist case, S imagines Is motorist T may or may not have; or S may simulate T's simulations of S. Simulation may operate off-line on either end: S may have T's actual sensory input, but disengage the efferent circuits, or else S would jump up, say, and perform the locutions that constitute R's warning to B, as some do when totally absorbed in sports events. Here, imaginary input leads to actual motor output. In these and other ways, one may simulate one's own and others' future states and also the future reactive attitudes to the future selves appearing in such deliberations. Combining several I/O mechanisms on-/off-line forms a complex intentional architecture – a rich hierarchy of on-/off-line elements. Those who can take their I/O causal mechanisms off-line have greater control over their I/O causal relationships with the environment than those who cannot.

I review some arguments between simulation theory and the theory-theory by Gordon (1992a, 1992b, 1992c), S&N (1992, 1993), and Schiffer (1992), *just to illustrate/explicate the features of the off-line model.* Developmental evidence (Baron-Cohen and Cross, 1992) suggests a tie between simulation and nonintentional (mimicry, gaze-following and other 'dumb') processes; intentional processes may be finessed as later developments, with an error theory for the intentional *language*, but no eliminative version of simulation theory has been developed. Schiffer has developed a plausible *no*-theory-theory (1987, 1991); simulation theory is also such a theory. Simulation is plausibly just an early jump-start stage in a developing *tacit* folk theory of mind. It is intuitive: We simulate first, and we form theories later; in the more primitive developmental stages, primate and human infant mimicry precede theorizing.

Simulation theory accounts for our tendency to ascribe mental states to others by our ability to simulate. Note that Gordon rejects the 'model' model of simulation (1992a, p.11),

in which we use ourselves as models of others. Gordon's move avoids objections based on the logic of model-based inference, by analogy with models of airplanes in wind tunnels (1992a, p.27).

> One general objection here is that the simulator himself must have a theory because ... to project one's mental states ... one first has to discriminate them, and this requires that one already have .... a theory. (1992a, p. 26)

Quoting Paul Churchland regarding a model of the universe, Gordon adds,

> Even if my miniature unfailingly provided accurate simulations of the outcomes of real physical processes, I would still be no further ahead on the business of explaining the behavior of the real world. In fact, I would then have two universes, both in need of explanation. (1992a, p. 26; citation omitted)

Gordon replies, "a manipulable model can be used to model counterfactual conditions" without a theory; to test explanations, we vary conditions without a theory (1992a, p. 27).

Another version of the 'model' model involves 'empathetic understanding' or *verstehen* (Gordon, 1992a, p. 31). Gordon rejects "putting oneself in the other's place" because it assumes one is not already in the other's place (1992a, p. 13). Rather than take a third-personal view, when asked to put oneself in another's shoes, "what is implied is that you shouldn't just project your own situation and psychology on the other" (1992a, p. 13). He calls projection without "patching up" (making the relevant adjustments) for differences between self/other "total projection", the default setting for simulation. The reason hiker B is not amazed by A is that "we are constantly projecting [onto others] our own beliefs about the environment" (1992a, p.16).

Another attractive feature of simulation theory is its ability to account for the *felt primacy* of simulations in contrast to the *aloof computational character* of beliefs/desires under the theory-theory. Suppose T is being approached by a bear, and S sees this and wants to understand T. On the theory-theory, S supposes T has certain attitudes and uses attitude-

mentioning laws which, in Fodorian practical syllogism form, yield the *proposition* that T should do something to avoid danger. On simulation theory, in simulating T, S gets simulated *fear*.

> Probably the most important use of this capacity ... is to reveal, not *what* the other perceives, but *how* he perceives it. For example, the grizzly that is approaching, not you, but your friend, who is some distance away, must be seen as an approaching grizzly: that is, one you might describe as '*approaching me*'. (Gordon, 1992a, p.19; emphases added)

Successful simulation uses "patching up" to account for differences between S and T. Without adopting the belief that the bear is *approaching me*, S isn't inputting T's belief.

The method of patching up handles "mother-in-law" objections that conflate elements of explanation and justification. As Gordon, paraphrasing Perner, says,

> Assume you learn that your colleague's mother-in-law has just died. How does he feel? It will not do to imagine that your mother-in-law has just died .... Because your relationship ... may be quite different from your colleague's .... Your simulation must be informed by some 'theory' about which ... relationships are ... relevant. If you love your mother-in-law but your colleague hated his, then your simulation will be more accurate if you imagine the death of one of your foes. (1992a, pp. 21-22; citation omitted)

But in such a case, we would ask, "Were you close to her?" We know from *experience* – not from a *theory* – that family role does not infallibly predict a standard relationship. It is circular if the objection is that "it must be a theory". The problem is that we lack the *information* needed to grasp the *specific* relationship, but inadequate information is an *epistemic* issue confronting *both* theories. Whether simulation theory or the theory-theory better explain this is irrelevant to our ends; again, my review of elements of these debates is *just to illustrate off-line elements.*

Let the epistemologist worry about justifying information derived from brute simulation mechanisms. Principles of *epistemic justification* are irrelevant to identifying *psychological* processes, such as "*How do we* predict, explain, ascribe intentional states?"

One may object that what we are concerned with is not just brute skills, but their *success*, and that invokes questions of justification, but the origins and justifications of our beliefs about others are distinct issues. We *assume* we're *all* getting it right, not that the debate hinges on who gets it right more often, tacit theorists or simulators – irrelevant to us anyway. S has the relevant information on T's relationship with his mother-in-law or not: If not, S will not get it right; if so, he will.

Developmental evidence does not favor the theory-theory, though drawn from a research paradigm designed to establish the theory-theory (Harris, 1992, p. 123). Let us turn to the evidence. Developmental studies were designed to support the "child's (tacit) theory of mind" used to grasp mental states. Though child *pretense* is central to the data itself, simulation *theory* is only a recent alternative. Much evidence was amassed from tests designed to age-isolate skills by means of false belief tasks.[35] In the initial model, the subject is shown two puppets, P1 and P2. P1 places candy in a box and leaves, P2 moves it to a concealed location, and P1 returns; the child is asked about P1's (false) belief, "Does P1 know where the candy is?", and about P2's (true) belief, "Does P2 know where the candy is?" But the *tests presuppose* ability to trade in the *language* of *belief* and *pretense projection* onto puppets!

Most 2- and 3-year olds err, thinking P1 thinks the candy is in the new location (though P1 would be ignorant of this). Some 4- and most 5-year olds rightly think P1 thinks the candy is in the old location. How should we interpret this? Does the younger child's inability and the older's ability indicate a development in a theory of mind, or in the ability to simulate? A shift from inability at age *n* to ability at *n+1* can *age*-isolate these skills, but

---

[35] Wimmer & Perner (1983); Baron-Cohen, Leslie & Frith (1985), S&N (1992), Perner and Howes (1992), and Gordon (1992a, 1992b, 1992c) and Goldman (1992).

cannot *type*-isolate them. Our theories are like rabbit *stage* and *part* theories under Quine's radical translation restrictions – *not* that there is *no fact to the matter*, but *these* test-design restrictions just aren't telling *as such.*

No contributor to this discussion has noted that the test's basic condition, the use of puppets, presupposes a child's understanding of *pretense*, thus of *belief* and *make-believe*, and their *projection* onto puppets. The child must grasp that puppets *perceive, believe*, etc., before any question of *false* belief even arises – before the tests are run. So, the 'research finding' that younger children are not yet versed in false belief is undermined *a priori.* This *basic* design flaw forestalls any theory selection – long before test conditions are made even more complex. The theory-theory cites age-isolated skills involving the complexity of *false* versus *mere* belief as evidence of mastery of folk theoretical posits such as belief, *but these are already presupposed.*

Simulation theory cites the same skills as steps in the developing ability to *patch up* total projection: The younger children get the false belief task wrong because they're employing total projection, e.g., "if I were him, I'd know the candy wasn't in the box, but in the cupboard, as I now know"; they fail to patch up, to adjust for relevant differences. The older children get it right because they're not just projecting, but also taking the other child's vantage point into consideration, and disengaging the relevant beliefs of their own; e.g., "if I were her, I'd still think, as she thinks, that the candy was in the box, because I wouldn't have seen it being moved to the cupboard". The test design obviously presupposes that children already have the knack of attributing beliefs. This presupposition can be given a theory-theory gloss, or a simulation gloss. *Ideally*, the *test design* itself shouldn't tell us which, and the interpretation of the results will rest on which gloss is adopted. So, at best, the experiment can decide nothing. But things are not ideal, especially for the theory-theory.

For, on little reflection, it is obvious that the test design requires subjects to use *pretense – make-believe*, a clear *hybrid* of *simulated belief*. We begin and end in a stalemate that supports my adoption of a hybrid view according to which *we at least sometimes simulate*, and that is enough for purposes of illustrating off-line activity without being vulnerable to the fallacious though popular objection properly called "attacking the illustration".

Even if developmental data showed children wielding a theory, adults may still be simulators or vice versa. Yet none of the data supply any reason to rule this out, nor to rule out the possibility of starting or finishing with a hybrid mechanism, or of going hybrid from start to finish. But a hybrid theory is *most* plausible, since mimicry *means nothing* if not *already* taken to be (Brentano-style) *intentional*. This is true even if later-developing, more mature, and thus greater understanding, which involves more fully transparent intentions, dovetails earlier and thus lesser mimicry, which involves opaque, pre-verbal, and thus pre-intentional psychological content. *Tests are not needed to see that we begin with mimicry and other brute simulational mechanisms,* as primate jump-start mechanisms that data-feed our growing understanding of mind. We accumulate knowledge both from simulations and from inductive and theoretical cogitation, confirmed by ubiquitous experience.

Gordon suggests that adults use simulational counterfactual reasoning to fill the gaps in *ceteris paribus* clauses in folk psychological generalizations (1992b, p. 91), which, if true, grounds the possibility of adult hybrid theories. Walter (2002) gives neuropsychological grounds for the idea that counterfactual reasoning is simulational. In talking about the emotional/neuro-physical basis of our pro-attitude identifications as a way to stop a regress problem facing hierarchical accounts like Frankfurt's, Walter states:

> There are concrete neurobiological hypotheses about this mechanism. The dorsolateral section of the frontal cortex is important in simulating future counterfactual situations. The ventromedial section admits mental sample actions

> into the evaluation circuit. This center joins emotional centers, the body, and its neuronal representation. (2002, p. 575)

Counterfactual simulations ground our deliberative entertainment of practical possibilities in a naturalistic mechanism, and support the chapter 2 claim that we need not access alternate worlds; for the functioning of this mechanism is an actual-world cognitive phenomenon. Thus, simulation's determinism-friendly causal/counterfactual properties help explain the phenomenology of autonomy. Deliberation may involve simulating our own possible desires/choices as I/Os; with these choices as further inputs, we simulate our/others' reactive attitudes towards us for outputs about the expected effects on our lives and characters. We juggle simulata, select from the overall set of resultant counterfactuals (e.g., "if I choose this, that will happen"), and then go on-line, and our doing so is *what causes that choice*.[36] As Levin says, the causation *moves through us* in just the right way; we are essential, not superfluous (1979). Though the functioning of the black-box mechanism is determined, it is in *control-generating*, rather than *control-losing*, ways – favored by evolution for maximal self-regulation – by the reflexive cognitive/conative centers of our brains.[37]

Let us turn to Schiffer's counterexample. The notions of justification in mother-in-law cases and Schiffer's case differ. Schiffer's involves normative rationality: An action is justified by its link with Agent's beliefs/desires. His counterexample (1992), which may not indicate his view, is that he may know some molester M has certain propositional attitudes, e.g., M wants to molest a certain subway passenger. On simulation theory, knowledge of

---

[36] Of course, the more reflective among us, like Hamlet, stay off-line longer and process simulata more and with greater iterative loopings than the less reflective among us.

[37] Indeterminists, with Mele (2002, 2003), think the generation of reasons-for-consideration is optimally indeterministic. That evolution designed this black box argues *against* more than a modicum of *pseudo-randomness* in a mechanism we do not fully control: Reasons-for-consideration are optimally determined by salience-stamping processes in long-term memory.

M's attitudes explains why M molested, for if Schiffer were to simulate M's attitudes, he should get the same choice to molest, and this would explain M's act as what anyone in M's shoes would have done. He sketches the sorts of attitudes that might cause M's choice, e.g., M believes M will not get caught, M strongly desires that sort of pleasure, etc. Schiffer objects that *his* M-simulation would not yield the molestation choice in Schiffer's own off-line processing of M's attitudes.

Why would Schiffer not arrive at the decision to molest, whereas M would? Schiffer is supposed to adopt whatever states M has when M molests, excluding his own intentions when they diverge. Schiffer's simulation should not diverge from M. That it does diverge suggests failure to properly simulate. Compare: A concentration camp prisoner asks a Nazi, "put yourself in my shoes"; the Nazi replies "If I were a Jew, I'd kill myself for Hitler". Schiffer and the Nazi fail to fully engage with the terms of the experiment.[38]

Schiffer's claiming to adopt all of M's intentional states but not be led by them to the same act entails that M's intentional states are causally insufficient for his act. If two agents with identical, widely-construed intentional states can be led to divergent decisions, we need not worry about determinism, for we would have contracausal freedom. But Schiffer's account of mental causation (1991) includes a causal-counterfactual component in which mental states are causally sufficient for actions. Indeed, all but eliminativists and acausalist libertarians, e.g., Ginet (2002), agree that intentional states are causally sufficient for intentional behavior, so this suggests a failure on Schiffer's part regarding faithfulness to the terms of the experiment.

Simulation theorists may cite a failure on Schiffer's part to "patch up for deficiencies

---

[38] On the dialectical use of thought-experiments, see Stalnaker (1984) and Sorensen (1992).

in total projection", i.e., correct S's errors in excluding S's intentions or including T's. Someone psychologically type-identical with M would do what M does under type-identical circumstances; so, if S correctly simulates T, S will arrive at type-identical output as T. Thus, "patching up" eliminates all such counterexamples and grounds an error theory about them: They are simply functions of only partial projection. *Normative* and *epistemic* notions of justification are distinct, though related *here*. *Normative* notions here are about what it is *rational* to do *given* M's I/Os. If left out of the *epistemic context* – the selection of input – needed for a *correct* simulation, we'll get the same sort of counterexamples as in standard *epistemic* (mother-in-law) cases. Schiffer's failure to input relevant data for M, including M's distorted behavioral norms, is a case in point.

The off-line theory may avail itself of the intellectual generosity of simulation theory's formidable critics. S&N (1993) reply to criticisms leveled against their earlier critique of simulation theory (1992), though they have backed away from the adversarial enthusiasm which marked their earlier paper. They contend that while simulation theory and the theory-theory seem to be the only two games in town, it is *an empirical question* which one better accounts for our folk psychological skills or whether both account equally well for separate skills. As they generously put it in their conclusion,

> [We distinguished] different versions of the simulation theory.... [These] distinctions are both more interesting and more valuable than the various polemical points we've tried to score. For they indicate that the dispute between theory-theorists and simulation theorists is much more complex than has hitherto been recognized. It may well turn out that some of our folk psychological skills are indeed subserved by simulation processes, while others are subserved by processes that exploit a tacit theory. Whether this is the case, and, if so, which skills rely on which processes, are matters that can only be settled by doing the appropriate experiments. It looks like there is lots of work to be done before we have a good understanding of how people go about attributing mental states to each other and predicting each other's behavior. (1993, pp. 28-29.)

Their concluding perspective is all we need for our weak, off-line- illustrative purposes.

There are neural reasons for thinking counterfactual/deliberative reasoning involves simulation (Walter, 2002). This grounds the *empirical – non-speculative – character* of the off-line theory. That simulation subserves *some* folk skills, even if theorizing does also, is the basic idea of the hybrid theory. I doubt anyone familiar with primate and human mimicry and gaze-following would deny that simulation is an evolutionary endowment of higher primates that humans use in developmental stages. Simulations may be instinctive mechanisms that jump-start processes leading to a theory of mind that results from reflection on its results over time. This innocuous hybrid theory is enough to support the off-line theory claim that we can take our cognitive/conative systems off-line, which is all that is needed to empirically ground, motivate, and naturalize the off-line theory.

The more important feature illustrated in simulation that supports autonomy is its off-line metamental component – part of mind operates on other parts of mind. This off-line feature was vividly illustrated in connection with simulation, where its features are more transparently visible, but it is attributable to metamentality in its own right, rendering the causal/functional 'off-line' concept fully independent of simulation. Two other ways to illustrate causal/functional features of features of off-line/metamental states and also the independence of 'off-line' and 'simulation' are by identifying other off-line/metamental behavior that is not simulational.

Some off-line/metamental behaviors not explained by simulation theory are *first-personally*-activated *mind-altering* techniques like meditation and sensory deprivation, or *third-personally*-activated *behavior-altering* techniques like Classical/Operant conditioning, brainwashing, or hypnosis. Others are *dreaming and language, and possibly memory. The mammalian brain is an off-line brain*, since it *dreams* (Winson 1997). These metamental functions link to form an off-line I/O manifold.

I define a *yogi* as a practitioner of an Asian *meditation* discipline. The yogi engages in off-line behaviors that do not require simulation, though some techniques may. The yogi's eyes are closed, taking the primary orienting-element of the *sensory nervous system* (visual perception, our primary mode of exteroception) *off-line*; his body is "locked" in a motionless posture, taking elements of the *motor nervous system off-line*. The yogi slows his heart/respiratory rates by artificial breathing rhythms, taking the cardiopulmonary system off-line from its sensorimotor-cued oxygenation processes. Attention is disengaged from its usual exteroceptive and proprioceptive links and directed to a mental target, such as the flow of thoughts. As these fade, brain waves slow down, and consciousness is taken off-line from its usual cognitive and conative environmental links and redirected into a brainwave state that is type-identical to a dream state.

Yogis enter these brainwave states while fully alert and awake, exhibiting the alpha, delta, and even theta brainwave state patterns associated with just falling asleep, rapid-eye-movement (REM) dreaming, and dreamless sleep, respectively. Yogis coordinate off-line behaviors under complex methodologies, producing off-line trance-like states whose corollaries have been empirically detected.[39] The yogis' off-line states provide a dramatic case of the simple off-line states posited for ordinary agency by the off-line theory, but both have freedom properties in common as a function of their off-line/metamental operations.

Let us next consider Classical or Pavlovian behavioral conditioning.

> In classical conditioning, one stimulus is associated with a second stimulus so that the first one comes to *stand for*, predict, or *signify* the second. The important result of this association is that the first stimulus is then able to elicit the same response as the second stimulus. (Olton and Noonberg, 1980, p. 35; emphases added)

---

[39] See Benson (1996), a rich source of additional citations to empirical evidence in favor of these phenomena. *Cf.* Stelmach (1976), Schultz (1965), and Madow and Snow (1970).

"Stand for" and "signify" suggest semantic relations, though more at *proxy*, which need not be *linguistic*. But language use parallels conditioning: Words substitute for things, and parallel the way stimuli are replaced by stimuli in conditioning. Language use and conditioning both involve substitutions, and these involve off-line elements.

Language use thus involves off-line interaction with linguistic substitutes (simulations) for things. By contrast, pre-linguistic species only interact on-line with non-linguistic things. Deliberators' interactions with the environment are mediated by linguistic simulata, thus at a further causal remove from the sort of lower-order, on-line causal engagement with the environment that is typified by the immediately-determined behavior of pre-linguistic animals.

Here is how experiments are structured in Classical conditioning.

> At the start of an experiment one of the two stimuli elicits an obvious response. Because this stimulus-response relationship was not developed through conditioning (at least in this experiment), this stimulus is called an *unconditioned stimulus* (UCS) and the response it produces is called an *unconditioned response* (UCR). The second stimulus does not produce a response, or at least not one similar to that produced by the UCS. If this second stimulus is always presented just prior to the UCS, it will eventually produce a response very similar to the UCR. The new stimulus is called the *conditioned stimulus* (CS) because it attained the ability to produce the *conditioned response* (CR) through conditioning. (Olton and Noonberg, 1980, p. 35; emphases and acronyms in original)

I adopt Olton and Noonberg's standard terms. Note that the UCR and CR are identical.

> [C]onsider [Pavlov's] procedure .... A tube was placed in the dog's mouth to measure ... saliva. [M]eat powder was presented to the dog for a few seconds [to see and smell]. In response..., the dog salivated, and the saliva was collected in the tube. Then a second stimulus, a bell, was presented by itself. [W]henever the bell was sounded, food powder was presented immediately.... After a few pairings..., only the bell was presented; some salivation occurred, but not as much.... The bell and food powder were presented together for more trials and then just the bell again. As the ... trials increased, the ... salivation in response to the bell alone increased until the bell produced almost as much salivation ... as the food powder had.... As a result..., the bell became able to produce a response that initially was elicited only by the food powder. (*Id.*)

In Skinnerian (Operant) conditioning, by contrast, instead of *UCRs* (from the autonomic nervous system), *voluntary* or "skeletal" Rs are manipulated by positive (feedback) reinforcements. So, a pigeon's voluntary pecking at a button is "rewarded" by the feedback of a food pellet, reinforcing that pecking. Research manipulating Rs led to the discovery that feedback mechanisms may link operant (voluntary) and classical (autonomic nervous system) conditioning, as in biofeedback.

These techniques dissociate/re-associate S/R pairs. UCS/UCR pairs may be viewed as 1st-order cause-effect pairs, for they are the unconditioned modes of *unmediated* causal interaction between organism/environment; they are on-line behaviors in that the environment-to-organism interaction is unmediated, the organismic R is a UCR, and the organism has *no control* over its R (no elbow room between S and R). Thus, conditioning takes S/R causal pairs, alters them (off-line relative to ordinary functioning), and puts them back in new concatenations.

The failure of behaviorism is eclipsed by the success of conditioning, where modification techniques target phobias, compulsions, and other forms of paradigmatically unfree behavior often used as litmus tests for theories of freedom.[40] These techniques take apart S/R pairs, in effect, and re-pair them in new combinations. The rearrangement of S/R pairs under *artificial* conditions involves an off-line element, for the original S/R interaction mechanisms – usually engaged in unmediated, 1st-order causal operations – are rerouted off-line. There is normally an unmediated stream of S-causes flowing from the environment into the organism and immediately back from the organism as R-effects into the environment, which unbroken causal stream completely controls the organism. On a hydraulic model,

---

[40] Ainslie (2001) uses behaviorism as part of "the data" for a theory of the will, since its research data is empirically rich, and theology, since its folk data is phenomenologically rich.

conditioning closes causal flow valves, disconnects pipes, installs new plumbing, reconnects the organism to the causal flow in a control-generating way, and then opens the flow valve. Therapeutic conditioning, with or without therapist-assistance, increases self-control, as do K's pills, e.g., arachnophobics use techniques to alter and thus control their spider S/R pairs. Deprogramming, hypnosis, and other forms of conditioning may also involve off-line elements, though illustrations from conditioning are less vivid than those from simulation or meditation practices. Again, 'off-line' is a *relative* concept.

In Classical conditioning, innate Rs (salivation) are manipulated by introducing *artificial* associations in artificial environments, producing new causal (S/R) pairings. The organism's I/O pairs are exposed to artificial conditions, taken apart, and concatenated in novel pairs. The I/O-rewiring is akin to S's running S's I/O mechanism off-line by introducing T's inputs, yielding T's outputs, for S's inputs are set aside and replaced by T's inputs, yielding outputs that are novel for S. This is certainly not as clear a case of taking a mechanism off-line as in the case of S simulating T, but there is sufficient functional resemblance to warrant the application of the on-line/off-line functional heuristic to the conditioning model.

In Operant conditioning, *voluntary* Rs are reinforced. Biofeedback links both conditioning forms, and converges upon the same phenomena the yogis have manipulated for millennia. That these off-line behaviors converge confirms the off-line theory. One link is the inverse relation between the degree to which a behavior is conditioned and the degree to which it is likely to be free. Spontaneous expressions are not highly conditioned and habitual behaviors are, but both may be controllable by Agent, i.e., changed if they do not comport with his higher-order intentions. Rather, the claim contrasts behaviors one is highly conscious of with volitions that operate without full awareness. Intuitively, the more

conscious an act is, the more it is intended, and vice versa; *cf.* somnambulism, an interesting case of unconscious but possibly intentional behavior that is intuitively unfree and amoral. *Consciousness of what moves you is the key to liberation* – this is also the method and goal of Buddhism (Keown, 1996).

Another parallel is the relationship between long-term meditation practice and how much control one has over one's emotions, how conscious one is of the causes of one's actions, and how conscious one is of one's own stream of consciousness. The long-term meditator self-effectuates behavior modification. Empirical findings support these claims (Goleman, 2003). Increased consciousness of one's mental states and actions is metacognitive and enables their increased control; intuitively, these are increases in autonomous functioning, i.e., freedom.

The knowledge of how to control future experiences is contained in countless conditional imperatives, e.g., "if you want x, then you ought to do y", so increasing knowledge of these conditionals can increase autonomy. Simulation of desires and their outputs can generate knowledge of conditional imperatives from the simulated I/O pairs. Suppose one simulates a novel desire together with related beliefs about its associated behavior, and then certain output decisions occur that diverge from one's own behavioral norms. A conditional imperative may be formed from this simulation experience, e.g., "if I want to become a divergent sort of person x, then I ought to do action y". Thus, simulation can yield knowledge about self-cultivation.

Similar insights inform doctrinal lore about wholesome (and unwholesome) mental states that may be cultivated (or weakened) in Buddhist meditation. The thoughts, emotions, and mental states that arise freely may be attended to with aversion, attraction, or neutrally; we also may step back and witness our natural aversion, attraction, or neutrality, and identify

with the witnessing stance, thereby inculcating some detachment from the intentional currents that normally move us to act. Wisdom about what mental states lead to what virtues/vices results from this off-line practice, not unlike the simulational one just described.

A meta-level insight from the results of many such simulations might be *to understand is to forgive*, since in simulating the causes of others' characters/actions, we see that we would be and act the same way, were we in their shoes. This sort of insight can defuse many reactive attitudes that rest on the assumption that we are better than others, suggested by Schiffer's assumption that he wouldn't get the same output as molester M. Thus, a conditional imperative drawn from this is that if one wants to be more compassionate, he ought to look more deeply into the impersonal causes of behavior and character. A similar insight arises from meditation practice. This doesn't threaten the reactive-attitude-web that constitutes our way of life, for to the extent one approaches maximal autonomy, he is an appropriate target for reactive attitudes.

Some think meditation entails an impossibility, i.e., pure introspection, but this is not so. To understand the possibility of meditation without introspection, consider an *interventionist* theory of observation: The mere act of observation affects the nature of the thing observed (Hacking, 1983). This was one reason Ryle thought introspection was impossible, since one cannot observe rage without interfering with it. But Buddhist meditation aims to *change* the mind *through* introspection.

The Rylean should accept meditation as a way of affecting one's mental states – "interventionist introspection" – while disallowing the possibility of passively introspecting them. Indeed, Freud, Chomsky, Fodor, and most other contemporary philosophers of mind

generally deny that consciousness is identical with intentionality.[41] The Rylean fact that the mere act of observing one's mental states tends to alter them constitutes evidence for the difference between consciousness/intentionality, but also for freedom through cultivation of awareness. The off-line theory claims that metacognitive awareness leads to freedom, and thus fits nicely with the Buddhist theory of freedom (Keown, 1996).

Nozick pumps intuition against the idea that there is any merit to meditation techniques and their associated ineffable states, but his attack will prove useful to the off-line theory. Let us call his argument the "soundless stereo" argument.

> Consider a phonograph system as an apparatus of experience.... Now let us do the equivalent of quieting thoughts, namely, removing the record, perhaps also turning off the speakers and the turntable. When only the amplifier is on ..., what is the experience like? We do not know; perhaps infinite.... It would be a mistake to think there is an unusual reality being encountered.... None of the literature I know describes what ... the procedure would produce in the absence of an unusual reality or self, so we don't know whether [it] is a revelation of an unusual reality or self, or instead an artifact of an unusual procedure of experiencing wherein most but not all functions are damped down. (Will this debunking explanation have more difficulty in explaining the surprising and often momentous changes in the people who have the experiences?) (1981, pp. 158-9)

Nozick's intuition pump critiques the yogi's *interpreted* experience, its metaphysics, not the assumption that the yogi can disengage conscious awareness from its ordinary functioning – take it entirely off-line. Thus, Nozick's critical analogy is enough for our purposes.

In Buddhist/Yogic philosophy, meditation is the path to liberation, construed as freedom from deterministic conative chains (*karma*). The Buddhist doctrine of "dependent origination" is a form of determinism; it says that *everything that is* emerges from, and is

---

[41] Fodor (1992). Some philosophers of mind favor sophisticated versions of that view; see, e.g., Searle (1992), chapters 7 and 8. Searle's view is not that consciousness and intentionality are actually identical, but rather that S is an intentional state *iff* S is in principle *accessible to* consciousness (the "connection principle"). Chalmers comes close to the connection principle in his discussion of the bridge law relating consciousness and its powers (1996). Searle's view is one of the closest to a full identity theory (for intentionality and consciousness) that I am aware of in the recent literature, but even his view leaves just enough of a gap between the two to allow for the possibility of interactionist introspection.

dependent on, *everything that was* according to essential (karmic, causal) connections. *Theravada* Buddhists practice *vipassana*, "insight" or "mindfulness", which suggests *freedom through insight.* Mindfulness is alert witnessing of what I describe as the 1st-order contents of consciousness. In basic practice, one first develops some concentration by focusing on one object of consciousness (Goleman 1988). After proficiency in "one-pointedness" is developed, one practices attentiveness to one's entire experience with no attempt to direct awareness, but rather to track one's conscious states wherever they lead without interference or control. *Zen* practitioners extricate awareness from the 1st-order stream of occurrent thoughts, emotions, sensations, by *simply sitting still and observing the stream of 1st-order mental contents.* This can create enough distance between the active (content-generating) and the passive (observing) components, though the 'passive' component interferes (Ryle, Hacking), which interference is what tames the mind and prevents the content-generating component from overcoming the 'passive' component.

By the practice of motionlessly witnessing the spontaneous arousal and cessation of mental states without interfering, but just attending carefully to the emergence of *intentions*, one develops an awareness of the *spontaneous* and thus *impersonal* character of their emergence, and thereby *gains greater control over the otherwise pre-conscious levers that guide our acts.* This is the *essence* of the Buddhist theory of freedom, akin to Freudian and other forms of insight-through-awareness and liberation-through-insight psychologies.[42]

---

[42] Locke (1690) remarks about *stepping back* in awareness to gain control over intentions; Hampshire sees awareness of the causes that move one as the key (1975). It has been suggested that Spinoza has a stronger view, that consciousness alone is free, though his intuition pump about the conscious rock strongly suggests illusionism. See Richard Taylor (1983) on Spinoza's claim about the rock. Krishnamurti (1962) and the *Zen* tradition advocate choiceless (non-contracted) awareness, as did Perls (1947), who also called his *gestalt* psychotherapy "*concentration* therapy". For a functional analysis of all major forms of meditation, see Goleman (1988); see also Goleman (2003), which narrates a convention on controlling emotion, involving Owen Flanagan, the Dalai Lama, and other interesting figures.

Mind seeing mind – mindfulness – is the key to *vipassana*. Awareness of psychological state x is thought to be the best route to insight about x. *Vipassana*, *zen*, and other forms of meditation have the common function of training attention by tracking the meditator's mental states. Proprioceptive tracking of kinesthetic states, which we share with other animals, need not be intentional, though metacognitive awareness of such states constitutes a hybrid kinesthetic-intentional state, and so is intentional.

We need not sort out all meditation-tracked phenomena, and just refer to them as mental states, though some may be intentional only insofar as they are mental (say, sensory) states of which one is mindful, since at least metacognitive awareness of (presumably non-intentional) sensations renders the meta-component intentional. Conative intentions are 'intentional' in Brentano's sense, in that they have an *aboutness* to them, as do thoughts, sentences, or other meaningful items that are *about* things.[43] But they are also 'intentional' in the teleological sense that they involve intent, motive, purpose. We may say that these forms of meditation are *cognition-and/or-intention-tracking*, and thus metacognitive.

We may compare *cognition-* or *conation*-tracking with Nozick's *truth*-tracking as a condition on knowledge (1981, pp. 172-78), *value*-tracking as a condition on goodness (pp. 317-31), and Wolf's notion of *reasons*-tracking as a condition on responsibility (1990).

---

[43] There are penumbral edges to intentionality. Imagine an infant that experiences heat, sound, etc., but with no interpretation of their source, or that they are *about* anything. Yet, they are impressions in consciousness, however opaque. After many same/different impressions are registered, pattern recognition occurs; later, interpretation is activated, and the infant senses the impressions are *of* something, some unknown x. Mothers tacitly capitalize on these processes by reproducing repeated word/object associations, but this alone couldn't transmute purely non-intentional impressions into intentional ones, were it not for the infant already being in a position to seek meaning or aboutness in these patterns. A way to conceive the difficulty clearly would be to try to program a computer to receive photo-impressions without pattern-recognition capabilities; if we add pattern-recognition, we solve the first problem, but what do we need to add to encode the ability to interpret them with the sort of understanding Searle alludes to in his famous Chinese room case? What is added to pattern-recognition so that an infant at a certain point knows its mother's word-noises have meaning? Simulation theory can explain this: Innate primatological simulation mechanisms, e.g., gaze-following and mimicry, set in motion redintegrative processes that bring intentions to life experientially as an infant matures. I apply this line of thought to the Gricean model of communication in chapter 6.

Wolf's and Nozick's notions are questionable, since truth and value are abstracta that may lack ontological substance, and they cannot literally be the object of inspection, but intentions certainly exist, and we can concretely inspect them. The parallels nonetheless suggest the utility of cognition- and conation-tracking (hereafter, "intention-tracking") as a condition on self-control and freedom from the bonds of (habituated, conditioned) self, for without knowing one's intentions intimately, one cannot help but be pushed/pulled by them.[44] Recall K's emotional incontinence and her pills.

One-pointedness attention-training may be functionally defined as *attention-* and *intention-training*, for it is designed to train conscious attention, of what one dwells on, thinks about, attends to, is aware of, etc., and these are all objects of one's interests, intentions, etc. The instrumental utility of the notion of intention-training ought to be clear: Attention-training fosters intention-tracking, and vice versa, and leads to intention-training (conative control). Together, intention-training and -tracking contribute to the formation of higher-order mental states involving an observer (tracker) and controller (trainer) of lower-order mental states: The 2nd-order tracker may attach a pro- or con- 2nd-order label, say, "desirable" or "undesirable", to any spontaneously-arising 1st-order state. Buddha, meet Frankfurt's hierarchy!

In fact, in canonical Buddhist psychology, mental states are wholesome/unwholesome insofar as they tend toward the maintenance of the meditative state (Goleman, 1988). Aristotle and his followers, e.g., MacIntyre, ground ethics in human nature; Buddha grounds ethics in soteriological utility, rendering it a pragmatic system of hypothetical imperatives of the form, *if one wishes to increase freedom, one ought to*

---

[44] See Hampshire (1975), and also (1987) (on Spinoza) for a similar view.

cultivate certain mental states and avoid cultivating others, i.e., those that foster meditative equilibrium or disequilibrium, respectively. Both avoid the naturalistic fallacy. Buddhist soterial ethics is thus a variant on ethics grounded in self-interest, geared entirely toward increasing autonomy.

Buddhism defines the virtues as states tending to sagacious growth through meditation; as Ryle said, rage is not amenable to introspection, so the theory defines rage as unskillful, poor *techne*. While I have no burden of formulating a moral theory, these definitions allow highly pragmatic value rankings of mental states and tendencies (virtues/vices) in terms of their propensities toward greater/lesser autonomy. This is a form of causal-knowledge philosophy.

Second-order intention-training may use one-pointedness to focus on and select or reject an intention, thereby making it occurrent or non-occurrent, thereby further contributing to hierarchical self-design. The result is a self-conscious, -designing, -controlling intentional hierarchy managed by higher-order intention-monitoring and -training functions, a top-down, Platonic, reason-over-desire, as opposed to a Thrasymachean, bottom-up, Humean structure, under which reason is (and ought to be) controlled by lower-order (unconscious) intentions.[45]

Ryle/Hacking interventionist accounts of observation suggest that observing one's mental event changes its nature; so, intention-training techniques alter content. This insight *implies* the possibility of self-control via awareness, which alters its mental contents. These higher-order mental functions work well together, as well as with the hierarchical intentional psychology in Frankfurt's account of freedom and personhood (1971).

---

[45] This pro-Platonic view goes against Hume's view of reason as mere slave to the passions, but without denying the necessity of passion for the possibility of action. Buddhists insist there can be no action, not even the movement toward enlightenment, without passion (the karmically conditioned – hence tainted, but "golden" or "noble" – desire to be rid of suffering).

Meditation typically involves several means of going off-line: shutting down the senses (e.g., closing the eyes), locking the body in a nonmoving posture, and withdrawal from the ordinary I/O causal interaction stream. This automatically takes the practitioner multiply off-line from her ordinary 1st-order causal involvement in the world. Higher-order mental states are at a causal *remove* from those of unfree creatures and so freer by virtue of this *distance*. Additionally, simulation occurs in *some* meditation practices, e.g., visualization, and accounts for an additional increase in freedom. (That *not all* meditation is simulational supports the independence thesis.)

There is ample empirical evidence that meditation techniques can be *control-enhancing*.[46] When successful, they lead to heightened awareness of and control over the forces that propel practitioners to action. Higher-order mental states are automatically at a causal remove from the unmediated cause-effect sequencing characteristic of animals, young children, and the deranged. Thus, meditative states increase freedom to the extent they lead to a vantage point – in terms of distance or dissociation – functionally removed from the characteristically unfree causal chain typified by unmediated S/R relations at the 1st-order on-line level of animal causation.

This is a theorem, and its consequences are predictions, of the off-line theory: We should expect these higher-order off-line states to be correlated with increased self-regulation.[47] That they are correlated is empirical confirmation of the theory. Thus, in virtue of the causal role of off-line mechanisms in consciousness-altering techniques, off-line mechanisms generally lead to added freedom. Locke and Spinoza, meet Buddha!

---

[46] See Benson (1996) for an extensive explanatory bibliography on these studies.

[47] See Mele (1995) for an analysis of the development from self-control to autonomy.

The root skill in meditation is concentration. This is developed by detaching *the part* of awareness *normally* focused on and identified with the 'stream of consciousness' – thoughts, desires, images, sensations, etc. – and redirecting it to a "primary object", target, or focal point. There is a facial tension in the appearance of awareness in both the role of subject and object of consciousness, but only a facial tension. For if I *notice* where my awareness *is directed*, my awareness is in one sense – that of my noticing – *subject*, but in another sense – that which I notice – *object*. The objection entails meta-awareness, which fits nicely with this analysis, for then meta-awareness, as in the meditative mind, is the most liberating among all the other meta-abilities: A metamental power to consciously direct lower-order cognitive/conative processes.

Being carried off by the push and pull of one's stream of (1st-order) mental contents, and drawing one's focus back to the primary target, develops concentration power – a (2nd-order) volitional power (to have some control over what objects of attention one will have) – with less effort over time.[48] The shift from *being led about by whatever mental contents occur* to *using one's volition to direct one's mental contents themselves* is like the shift from involuntary to voluntary breathing. This increases freedom, like the shift from *non-prehensile* to *prehensile* digits, which led to our ability to restructure our environment. When we learn to maximally control our minds, environments, and genetics, we will control what controlled us.

In meditation, pulling in the reins on 1st-order functioning is a 2nd-order mental exercise, and each repetition, as in resistance training, is freedom-training; practice makes

---

[48] Csikszentmihalyi (1991) cites evidence for lesser neural activation (in brain centers correlated with attention) in the use of learned skills. Evolution favors organisms capable of attention training, for lesser neural work for skilled attention frees neural energy for greater monitoring and processing of peripheral environmental stimuli, a clearly adaptive ability.

perfect. One practices freedom while being pushed by one's 1st-order desires and, more particularly, pulling back against their currents. Since one is sitting cross-legged with eyes closed, not actually led by one's intentions to act while open-eyed in the external world, meditation is a *practice* behavior.

Practice sessions, drills, or rehearsals are off-line in the sense of not being the real thing; consider the difference between shadow-boxing and a real match. By practicing in this way, one extricates oneself from the ordinary stream of consciousness (by disengaging attention from the impinging cogitation, sensations, etc. and returning one's attention gaze to the intended target). This trains the ability to go off-line, not only from 1st-order cognitive/conative engagement with the external world, but also from higher-order cognitive/conative states. The simulation theorist may try to bag some of these phenomena, given their facial affinity, but the affinity is only facial.

There is a parallel between attentional and resistance training. Body size, strength, agility, and overall physical prowess improve with resistance training; likewise, general mental powers improve with meditation training: alertness, attention, concentration, visualization, focus, and self-control. The yogi's *brain* improves (Goleman, 2003, chapter 8), as we might expect, given its plasticity. Consider the differences between those who do no attention training, those who do some, and those who do a lot. These differences yield quantifiable differences in off-line powers. Those *who try harder to be free* can develop greater self-controlling powers than others. The implications for self-cultivation and responsibility for self ought to be clear.

There is a facial conflict between standard Western and Buddhist views on the relationship between self and freedom. On Frankfurt's view, personhood is a necessary condition of freedom. On the Buddhist view, though not on the other yogic or "Asian"

views, self is an illusory entity, an ignorance-based obstacle to freedom that vanishes before the liberated, enlightened mind. But a greater commonality is visible in the fact that the off-line and most Asian models agree on two soft determinist points: the inverse relationship between freedom and lower-order conative determinism, and the consistency between global determinism and freedom.

Buddhists view mind/self as a mere cluster of transient physical factors; *cf.* Hume, Ainslie (2001), Parfit (1986), and G. Strawson (1986) for no-self views of mind. On my account, self is an off-line construct with psychological, causal, and ontological significance. Self-creation involves identification with simulated projections about the sort of person one prefers to become: One develops 2nd-order pro-/con-attitudes towards the projections, and identifies one's character as one's 2nd-order preferences; the self emerges as a causal-power-possessing by-product of the hierarchical architecture of the off-line mind.[49] Thus, as the self matures, it takes on an integrated hierarchical structure with its own causal powers and thus its own life. Similar things may be said for autonomy. There are quantifiable and thus ontologically significant differences between the causal powers of the mind of the accomplished Zen master and the non-meditator, as between the rational deliberator and non-deliberating creatures.

Drawing some of these ideas together now, I outline a model of deliberation, decision, and action that extends simulation's off-line mechanisms. Simulation mechanisms operate by taking some cognitive/conative (I/O) system off-line, feeding it pretend inputs, and letting the system produce simulated outputs. Suppose simulator S disengages his

---

[49] For Perls (1947), the self is a homeostatic (self-regulating) *process* forming the system of identifications/alienations as the organism defines itself through its "gestalts" (figure/ground relationships) – personality functions. Frankfurt (1977) and Nozick (1981) say similar things. Perls says organisms are equilibrium- and growth-seeking, homeostatic, or self-regulating.

belief/desire/action system; it functions off-line. Were we unable to control on-/off-line system functions, as in the boxing example, we would have a difficult time adjusting beliefs and behavior in truth-sensitive ways. It follows that we can operate our I/O mechanisms off-line in ways *not* immediately belief-forming or behavior-producing, as they *are* when on-line. When on-line, as with animals that cannot go off-line, I/Os are immediately connected by on-line functioning of the system, despite an intervening nervous system that makes the behavior more complex than the simple Hobbesian transference of motion, say, from one billiard ball to another.

Thus, if an I/S is provided to a non-off-line-capable animal, say, a torch is flashed before the eyes of a snake, it will produce a certain behavioral O/R, say, it will spring backward and hiss. Though the I/O is mediated by the snake's nervous system, we distinguish between the snake's nearly immediate I/O and simulator S's highly mediated I/O.

The objection that both behaviors are ultimately S/R-explainable, ours being more complex, parallels the hard determinist's objection that both behaviors are ultimately determined. But this ignores the features of the complexity that render the distinction one with a difference *greater than mere complexity*. And that is that off-line reflexive awareness of how one's and others' I/Os operate within one's on-/off-line system generates *the possibility of elbow-room* selection from among I/Os and *control over them*, control developed through exercise. Whether the I/O for on-/off-line systems is still amenable on some level to a highly-complex *S/R-type* description is not at all obvious. For the behavior visible in off-line-capable organisms seems facially of a different *type* than that caused by simple S/R pathways, just as the micturation visible in continence-capable organisms seems of a different *type* than simple micturation. To insist that S/R-type causation *must* apply to all behaviors *a priori* is to beg the question here. Continent and incontinent behaviors are

both caused or determined, but only one is additionally controlled by Agent in observable, causal/functional terms which, while explainable in a lower-level vocabulary, remain causally distinct nonetheless.

S's simulation consists of a simulated input of T, and a simulated output. S may come to believe *on-line* as a result that T's output is reasonable in light of his input. This illustrates the connection between ability to *go off-line* and to *contro*l ourselves. Without S's on-line desire to understand T, S would not have engaged in off-line activity. This models how the initiating causes of off-line behavior may be on-line. As a result of the off-line activity, S wound up with an on-line belief about T. Thus, some causes of on-line behavior are off-line. Off-line behavior may cause other off-line behavior. Causation may move through the hierarchy in any direction.

This is no reason to suppose that on-/off-line operations are type-identical. Nor is this ground for the CON-type claim that since the causes of both off-/on-line behavior are ultimately prenatally on-line, none of the primary causes of human behavior are controlled by Agent. To the contrary, the difference between off-/on-line behavior – between causes primarily in/out of Agent – is of so great a degree that it warrants being treated as a *virtual* difference in kind. This claim, *that off-line or higher-order mental functions are a virtually-distinct kind of determinism-friendly agent causation*, but a subspecies of ordinary causation nonetheless, is an Archimedean point.

There are many arguments in the literature in support of the claim that mechanisms-based explanations are genuine forms of scientific explanation (e.g., Cummins, 1983), and simulation is a mechanism. F&R (1998) take their reasons-sensitivity mechanism, implicitly *qua mechanism*, to be compatible with determinism though quite capable of causal/counterfactual dispositional powers; they assume mechanistic phenomena are

intuitively deterministic. If that assumption is correct, then simulation *mechanisms* are equally determinism-friendly; they are more genuinely *mechanisms* than reasons-sensitivity is, for *reason* is a 'mechanism' only figuratively.

Our meta-ability to take mechanisms off-line is an evolutionary development,[50] peculiar to us in degree, but not in kind. The differences in degree between brute mimicry and deliberation are so great as to qualify as a *virtual* difference in kind. Our ability to take our mechanisms off-line from ordinary causal I/O operations in deliberating warrants the view that we author our actions endogenously, though the causes of off-line behavior are ultimately exogenous. Since authorship is central to free action and responsibility, this model grounds a wide range of intuitions about freedom and responsibility.

Since freedom involves the meta-ability to take I/O mechanisms off-line, and comes in degrees, the off-line theory grounds the intuitions that (1) the amount of freedom depends on the satisfaction of physical conditions; (2) freedom admits of degrees; (3) freedom is an evolutionary trait of higher mammals; and (4) young children and most animals are generally not free. It also suggests an empirical direction for the lines of demarcation.

The higher-order architecture of our minds supports an account of the difference between us and Sphex, Dennett's tropistic wasp. The meta-structure depicted by the off-line theory can help to fill in the details of what makes for the possibility of more "elbow room" in our case. The off-line theory accounts for our ability to be flexible rather than locked in mechanical sub-routines, making it reasonable to call us "Flex" rather than just complex Sphex. Agents need not go off-line, nor need they choose consciously, to act voluntarily or

---

[50] Why natural selection favors mimicry is revealing. Given the nomic relations between emotions, facial gestures and pre-verbal utterances (shrieks, etc.) (Baron-Cohen and Cross, 1992; Bogdan, 1997, 2003; Conniff, 2004; Reynolds, 1993; Tomasello and Call 1997), it follows that organisms that mimic have redintegrative access to other organisms' intentions. Such mind-reading creatures make for good predators and poor prey.

freely. But insofar as *we are skilled in doing so*, we need not react immediately to every S, for we are able to go off-line and create a good deal of elbow room between S/R, when appropriate. We can refrain from responding instinctually, habitually, or out of unconscious conditioning. When we can and ought to go off-line or refrain, but do not, we are responsible for our actions.

The elements of life describable from the intentional stance seem jeopardized by determinism: Since we're determined to choose what we do, rationality need not bear on our thought, choice, or action. But the off-line theory accounts for the relevant practices in terms of our abilities to engage any/all of our I/O systems in on-/off-line operations of our systems for purposes of interpreting or controlling behavior. On this model, we are the authors of our actions, inferences, etc. For we can disengage these mechanisms from the environment, and thus set in motion off-line processes under our causal control, even though both on- and off-line processes are ultimately caused on-line. This gives substance to Hume's contention that it's not the fact that things are caused that makes them free, but how they are caused.

One may worry about attributing intentionality here to animals. There is a difference: Though it involves mentality, animal mimicry need not be intentional. Thus, the off-line theory doesn't attribute it to mere mimicry-capable beings. It is reasonable to mark the difference in causal mediation between the *immediate* S/R pairing in on-line animal functioning and the *mediated* I/O relationship in simulating agents by calling the former a *1st-order causal function* and the latter a *metacausal function*; there may be more than one meta-level. The *causal/metacausal* distinction is more important than the off-/on-line one. All that matters is that we have metacausal abilities, and exercise them in deliberation.

The metacausal functioning of agents' I/O mechanisms constitutes only a mediated engagement with environment, and so is at *a causal remove* from the animal's full

engagement. Thus, off-line activity is aptly-described as disengaged from immediate causal interaction. Thus, we have the elements of a cognitive architecture – an off-line causal hierarchy – for the off-line theory: Insofar as agents' deliberative and decision-making mechanisms involve metacausal functions, they are at a causal remove from the on-line causal nexus that strips beasts of elbow room and causal authorship of their actions. Insofar as agents operate at higher levels of metacausal complexity, to that degree the *pertinent* causes of their behavior are properly located within *their* cognitive/conative architectures.

Since practical reasoning is metacausal, the reasons/causes dichotomy is replaced by a model that (1) grounds their distinctness, and (2) reconciles them because reasons are metacauses. The meta-level distinction for off-line mechanisms is causal, but enjoys the benefits of hierarchical theories per se, such as Frankfurt's meta-desire model (1971).

Frankfurt's model has been accused of being overly intellectual. Since freedom rests in his analysis on iterated intentions, it requires sophisticated intellectual development and places a top-heavy emphasis on higher- as opposed to lower-order mentality, or lofty as opposed to base motives (Stump, 1993). But the off-line theory is less vulnerable to the top-heavy or overly-intellectual objection, since the off-line mechanism is a brute, evolved-primate ability. His model is also open to an infinite regress (Watson 1975). For it rests on the tacit formula that *agent A's desire d is free at level L1 iff A approves of d at L2*. But the same formula for transmuting unfree d at L1 by viewing it from the next level up L2 may be applied to the L2 approval, which is viewable from a yet higher-up level L3 as unfree, and so on. There seems to be no principled reason to stop the regress up the hierarchy.

Frankfurt's attempt to stop the regress, by almost stipulating that Agent stops the regress by making a decisive 2nd-order *commitment* that he identifies with, is unconvincing. I said "almost stipulating" because Frankfurt claims that the case is analogous to cases in

which an arbitrary decision is superior to no decision, as in, say, the Buridan case. This seems both ad hoc and inapplicable here. But even if it is neither of these, it seems unconvincing because it leaves open the brute fact that at the very peak of a model that looks upward for freedom, there is none.

The off-line theory is immune to the regress objection, for its hierarchy is not generated by iteration of intentional features, but from *causal/functional* reflexivity. So, its hierarchy is purely causal, non-intellectual, and non-iterative. Thus, the off-line theory lacks a formula analogous to *d is free at L1 iff d is approved at L2* that would be subject to iteration. My theory is deeply concerned with the same meta-abilities as is Frankfurt's. I deploy a causal/functional analysis of autonomy, but Frankfurt explicitly eschews causal consideration, so our theories are distinct, and mine, being causal, categorically escape the regress problem his theory faces.

Metamental functions may operate off-line. Simulation and other I/O mechanisms may be taken off-line. It is a non-intellectual fact that cognitive-processing mechanisms function well only at lower levels of hierarchical organization, due to data-processing limitations. There is only so much complexity that can be cognitively managed in a mental operation, e.g., what A thinks that B thinks that C thinks, etc. The same is true of any off-line mechanism.

For Frankfurt (1971), L1 desire d is rendered free at L2 *iff* we form a pro-attitude d' about d at L2, but the pessimist can ask about the justification of d', generating a regress. But the detachment of on-/off-line connections is what realizes our self-regulating abilities, and it is at root a serendipitous biological process with natural processing limits that non-arbitrarily terminate its functions (Walter, 2002, p. 575); only a few iterations *can* occur. Frankfurt needs a principled way to stop the regress; given my model, I have no regress to stop.

Frankfurt does not say what makes metaintentions possible, or how we form them. Wantons are *stipulated* as those for whom no inter-order conflict is possible, and persons as those for whom it is possible. This is an incompleteness in Frankfurt's account. It is also a confusion that the heroin addict who *lacks* a 2nd-order intention about his addiction is a wanton, for he must possess the *ability* to form a 2nd-order intention about his addiction which he has not exercised. We can imagine one with a 2nd-order desire to be a heroin addict (say, borne of her days as a flower child, who romanticizes drug culture). More dubious is that the addict is also a wanton with respect to everything else, that she lacks *any* 2nd-order intentions about anything.

The off-line theory has no such problems. Simulation is one mental function on other mental states; there are others. This cognitive architecture allows us to form 2nd-order desires. This is achieved by the virtual stepping outside of the 1st-order causal stream when the system goes off-line and creates a wedge between the 1st-order desire-inputs and their usual action-outputs, and makes possible a division between mental states in which 2nd-order desires may be formed. The division enables one mental component to have a pro-/con-attitude toward another.

We may view our ability to dissociate from simulated output as the first stage in the development of pro-/con-attitudes – identification and alienation – towards our own intentions. Given the dependence on this ability of both Frankfurt's theory of the person (1971) and Perls' theory of the ego-function (1947), this ability may enable autogenous self-creation and thus responsibility for self. Natural selection, on our just-so story, *may* explain this mechanism.

Consider what might happen when the deliberation mechanism is taken off-line and used for simulation purposes. Suppose S's output from simulating target T's input is a 1st-

order desire L1(d*) to do action A, but S has a con-attitude L2(~d*) toward d*. S must be able to detach from L1(d*). If not, S might automatically adopt T's pro-attitude L1(d*), and, if stronger than S's con-attitude L2(~d*), S might actually do A. Dissociation from simulated desires generates higher-order pro- and con-attitudes toward the simulated intentional output or simulata within one's own psychology, and thus realizes *the general ability to form metaintentions*. This is the sort of just-so account of how metaintentions *may have* evolved that Frankfurt needs, but lacks, in my view.

This may be seen as a same-order desire conflict. The difference is that the system's *off-line* functioning safeguards against S's identification with the T-simulation. S knows the system is operating off-line on T's intentional I/O, and thus S does not identify with output L1(d*). This makes possible, indeed *requires*, identification with some mental contents and alienation from others. (Of course, pathologies, e.g., schizophrenia, may involve cross-wired errors in these and other mechanisms.) This model also applies to self-formation.

I spell out a model of self-development later, but sketch it here. Perls (1947) analyzes the personality as a *function* of our identifications/alienations – what Frankfurt claims is more crucial to our motives than their source, history, or cause – that emerges from the organism/environment contact boundary over time, individuating the dynamic self *process*. On the element of our model just adumbrated, when S develops the ability to trade in pro-/con-meta-attitudes about simulated intentions, he also develops the ability to form 2nd-order intentions of his own, and to identify with some of these over others. In so doing over time, S forms his own self.

This gradual process has stages, with its earliest stage giving birth to, say, a mini-self that sharpens its initial boundaries with the environment through its initial pro-/con-attitudes toward features in its environmental field, typically a function of its

genetically/environmentally-determined hedonic states, e.g., proprioception of pleasant stimuli associated with its mother/caretaker and unpleasant ones associated with the latter's absence. It does this until it becomes a larger/sharper self, which self then refines its boundaries more, becoming yet sharper, larger and more autonomous.

Of course, much of it is directly exogenously determined in the self's early stages. But once the self reaches the stage where reflexive awareness of its own operations occurs, these are then subject to the self's own reflexive self-sculpting, which is significantly endogenously determined, even if it is all ultimately exogenously determined. This is analogous to the exogenous determination of continent micturation, which remains importantly distinct – in terms of causal powers – from the merely directly exogenous determination of incontinent micturation.

To turn to the mechanics of alienation, Schiffer's molester M case (1992) illustrates this point. Suppose Schiffer did get output O that he should molest the nearby woman W now. Unless he could detach himself from O, he would molest W. Schiffer's claim should be that he can detach himself from O, and refrain from molesting W. If we can do so with motives we recognize as *not* our own, *it is a tiny step from there to ability to do this with motives that arise in us naturally*, once we have a reason to develop a con-attitude toward them. And herein lies the possible natural genesis of autonomy – the ability to redesign our own wills from the meta-level.

Thus, S can detach from O; he can prevent simulated I/Os from entering straightaway into their belief/desire/action "boxes". This generalizes to any I/O from any off-line mechanism, and accounts for our ability to generate, and distinguish between, intentions and meta-intentions of various types, which we may think of as possessing subscripts representing our identification with, endorsement of, or other form of value- or weighting-

embedding for, those intentions and meta-intentions. Consider where T's belief/desires lead to an off-line result S considers desirable, but S lacks some of T's inputs. As a result of seeing the effectiveness of T's 1st-order intentions (in terms of the actions they produce), though S lacks them, S develops 2nd-order pro-intentions toward T's 1st-order intentions.

Suppose T is a lawyer and S's boss, and responds to S's questions about how T became a lawyer. T believed he had what it takes intellectually, loved to study the law, and wanted more than anything else to be a lawyer. Suppose S is an undergraduate philosophy major who, prior to this, had no great love of the law, but looked admiringly upon the profession, believed he had the intellectual talent and wanted very much to be a lawyer. As a result of simulating T, S develops the 2nd-order desire to have the 1st-order love to study the law, since S believes T's 1st-order desire to study law is instrumental in success at lawyering. This result might lead S to study the law with an attitude of discovery, looking for ethical principles and other jurisprudential underpinnings which might help him learn to love the law. Thus, by simulation, S develops a 2nd-order intention lacking a 1st-order parallel, but which contributes to its development. (Here, simulation yields a meta-desire to develop a 1st-order desire S lacks, which *causes* the formation of the lower-order one over time.) The directions of these relationships between S and T, and between S and S himself (engaging any mental mechanism off-line), across hierarchical levels, opens a rich set of possibilities that includes all the tools needed for autonomy, and much more.

This ability to identify with or detach from 1st- and 2nd-order beliefs/desires is a predictable byproduct of the partitioning and distancing generated from the functioning of the off-line mechanism. This ability may *first* function only *interpersonally*, when S simulates third-personally described person T, but extends to all intra-/inter-personal combinations/directions. Here lies the inter-personal basis of our ability to detach from our

own 1st-order intentions, and identify with some over others as a result of off-line deliberation processes. This is our architectural model for self-regulation, i.e., autonomy, and it is an important advance beyond Frankfurt's formulaic (L1/L2) metaintentional mesh model. Ours also connects with the Wittgensteinian idea that an individual's inner life is largely parasitic on his relations to others.

This analysis suggests that autonomy may be a by-product of *social animals* for whom evolution has endowed mind-reading or "heteropsychological" instincts for mimicry, linguistic representation, and other mimetic means of redintegrative access to other minds. These instincts, inclinations, and skills are clearly favored for smart social animals.

I call "primate emotivism" the tendency of primates to express pro-/con-attitudes toward attractive/aversive social behavior through the instinctive expressions of teeth-displays, facial gestures, body language, shrieks, and other pre-linguistic forms of intentional communication (Bogdan, 1997, 2003; Byrne, 1995; da Waal, 1982, 1989; Gomez, 1990a, 1990b, 1991, 1994, 1996a, 1996b; Kummer, 1995; Povinelli, 1996; Povinelli and Eddy, 1996; Singer, 1994, pp. 57-112; Tomasello and Call, 1997). The ability to regulate one's behavioral tendencies by way of the ability to identify, then to identify with or dissociate from, socially-approved or disapproved behaviors is clearly one that is favored by sociobiological natural selection.

Our mimetic instincts may provide the basis for our high level of skill with these interpersonal exchanges. Goldman (1993) argues simulation is the primatological basis of empathy, since it enables us to experience each other's experiences from the inside, which likely trigger self-discovered insights about our commonality.

Simulation *may* also subserve language-comprehension/communication in a Gricean way: When I simulate sentence X, I experience mental state Y. If so, these and many related

factors operate within and largely define the parameters and contours of the developmental environment in which smart social animals like us are cultivated (Bogdan, 1997).

Recall Wittgenstein's point that a great deal of our intrapersonal psychological states are importantly heteropsychologically interdependent.[51] Our analysis identifies their many interconnections in ways that exceed the sort of connections visible under competing models.

Let us spell out some of the further mechanics of our analysis. Let D_n^m be the m^{th} desire of level n. Suppose off-line deliberation produces the result that a 1st-order desire input, D_1^1, has negative value (consequential or otherwise) and another 1st-order desire input D_1^2 has positive value. Together with the psychological wedge created by going off-line, this output ranking makes it possible for the results to be fed back into the (metaphorical) desire box, so that D_1^2 is now preferred over D_1^1. Perhaps the result is not fed into the desire box straightaway, but is first embedded in a belief sentence about the 2nd-order preferability of D_1^2 over D_1^1, and perhaps then registered in the desire box, or in a higher-order desire box, e.g., $B[D_2(D_1^2>D_1^1)]$.

As the lawyer case illustrates, having the 2nd-order desire D_2^1 for D_1^1 does not entail having the corresponding 1st-order desire D_1^1. Thus, the off-line model distinguishes 2nd-order desires and volitions, the latter of which are simply those 2nd-order desires that succeed in engaging their corresponding 1st-order desires (Frankfurt, 1971). 2nd-order volition is *made possible* on the off-line account by virtue of the meta-level wedge or psychological distancing that is generated by the off-line functioning of the simulation and other off-line deliberative-mechanisms. Identification and alienation are explicit and crucial in Frankfurt, but not what makes them possible. This is shown on the off-line model. Identification

---

[51] On the primatological origins of the social interdependence of psychological states, see Bogdan (2003), p. 18, Bowlby (1982), Dunn (1988), and Travarthen (1993).

makes it possible for a desire to become effective and alienation makes it possible for a desire to be restrained. The off-line architecture grounds the possibility of iterated desires, and also meta-volitions, the voluntary, restraint, and other autonomy abilities.

By taking their 1st-order desires off-line and putting their 2nd-order desires on-line, or by running their 1st-order desires through off-line deliberative processes in which their 2nd-order desires displace their 1st-order desires, agents can dissociate from 1st-order desires. Dissociation allows the psychological distance – elbow room – between desires and volitions-to-act needed to prevent unendorsed 1st-order desires from issuing in action. Only with conflict between levels of desires, Frankfurt argued (1971), is it possible for freedom *to be a problem*, for only under this scenario can Agent want to do one thing but do another, as with the resistant heroin addict;[52] the wanton is no more 'unfree' than a worm. Since off-line functioning provides for iterated volitions, and these account for the problem of freedom, the off-line model accounts for the problem of freedom, and thus for *akrasia*.

We can account for akratic action as an *occasional* hierarchical conflict in which some 1st-order desire not endorsed at the 2nd-order level is stronger than a 2nd-order endorsed 1st-order desire. One formal difference between unfreedom and *akrasia* is that *akrasia* is local/occasional, whereas unfreedom is global/constant. No animal lacking metadesires *can* act akratically, for mere 1st-order-desire or Buridan struggles cannot generate the *phenomenology* of *akrasia*, e.g., identification with a weaker desire, regret that the weaker desire lost, judgement that the winning desire should not have won, etc. The details of a satisfactory account of *akrasia* need to be spelled out, but it is plausible to hold that they may

---

[52] Frankfurt argues rocks don't suffer freedom's *privation*; thus, we only adjust the criteria for autonomy by beings for whom freedom *can be* a problem. Shatz ignores the contrapositive of this privation restriction, like calling rocks blind because *not sighted*; he objects that rocks satisfy the conditional criterion (*if* they wanted to do otherwise, they would) (1986). Surely the main trouble with this objection is that the antecedent is too fantastic to take seriously.

be developed along these lines.

The off-line analysis suggests a probable – mostly inchoate, but empirically and conceptually pregnant – relationship that may be seen in what I call the "distance principle":

> The greater the distance between a higher- and the 1st-order on-line state, the greater the *potential* degree of freedom attached to the higher-order state.

The greater the functional distance between the off-line, higher-order states characteristic of free agents and the virtually unmediated on-line I/O sequences of 1st-order states characteristic of plants, *Sphex*, and other unfree beings, the greater the potential freedom.

Distance between hierarchical levels may be measured in simple quantitative terms as the number of levels in a hierarchy, e.g., a distance of one between the first and second levels ("hierarchical distance 1" or "HD1"), of two between the first and third levels ("HD2"), etc. With this measure, the principle may be expressed formally as an ordinal scale ("Distance"):

Distance: $(x)(y)\{[(Mx\&My)\&(Hdx>Hdy)]\rightarrow(Fxy)\}$

where "Mx" means *x is a mental state component*, "HDx" means *x is at a certain hierarchical distance from the 1st-order on-line state*, and "Fxy" means *x has a greater degree of freedom than y*. While the value of the formula is obscure, since its predicates are highly complex, it is designed only to capture the inchoate intuition that factors such as the ones it mentions make it the case that HD2 is likely freer than HD1. The greater the distance between hierarchical levels, the greater the causal/functional break from the 1st-order causal stream. Each level, being at a causal remove, functions as a wedge or gate between Agent and the 1st-order causal nexus.

Distance may be measured spatio-temporally. Spatially, it can be the literal – spatial – distance between S/R. In phobia treatment's "stimulus hierarchy", the target stimulus S (i.e., the S that causes the greatest anxiety) is placed physically at the most distant point on

the cognitive horizon, and Ss that cause less anxiety are arranged so that the less anxiety associated with an S, the closer to the patient S is placed (Olton and Noonberg, 1980, p. 39). Alternately put, this is the distance between the aversive S when situated by a behavioral therapist and the negative R. Temporally, it can be the lag time between S/R, in two ways. One way parallels the spatial sense: The lag time taken to close the gap between patient and target S. The other measures lag time between S- and R-occurrences; one subject takes .5 seconds to respond to S, another takes .75 seconds to respond. These spatio-temporal distance principles may be sketched formally:

Spatial Distance: $(x)(y)(z)\{[(Sx\&Ry\&Rz)\&(Gxy>Gxz)]\rightarrow(Fyz)\}$

where "Sx" means *x is a stimulus*, "Ry&Rz" means *y&z are responses to S*, "Gxy" means *x is at a certain spatial distance from y*, and "Fyz" *means y has a greater degree of freedom than z.*

Temporal Distance: $(x)(y)(z)\{[(Sx\&Ry\&Rz)\&(Txy>Txz)]\rightarrow(Fyz)\}$

where "Sx" means *x is a stimulus*, "Ry&Rz" means *y&z are responses*, "Txy" means *x is at a certain temporal distance from y*, and "Fyz" means *y has a greater degree of freedom than z.*

These formalizations only seek to approximate the principles needed to render transparent the underlying inchoate intuitions here, but no need for their more specific development presses here. These distance principles coalesce. They may also be correlated as temporal and spatial distance are in the S-hierarchy of behavioral medicine: The farther away S is in space, the more time needed to close that spatial gap. Other off-line activities can operate in tandem, each contributing its own measure to the distancing equation. These coalescent processes may form a manifold of integrated off-line processes, different by virtue

of the *summing* of distance factors, and by virtue of novel kinds of higher-order functioning. Some concatenations may be more significant than others in terms of metacausal powers. If distance principles hold, the more complex the metastructural state, the more distance factors increase, and thus the more freedom.

While lag time helps but is no *guarantee* of freedom, LeDoux's remarks about the different temporal distances involved in the almost immediate brain processing of emotionally significant phenomena and those of emotionally neutral processes are relevant (1997). Conscious processing is wired by natural selection to take longer. Ainslie's (2001) major concern with "hyperbolic discounting" is the hyperbolic shape of the discounting curve – distinctive of Ainslie, not just the fact that the present is valued more highly than the future. That is, it is the variable rate at which the *temporally distant* future is discounted which is important. That the hyperbolic tendency is the default for most organisms is instructive: Temporal distance can work against us, when distal rewards are less salient. But this itself supports my distance principle: The distance between Ss and Rs *decreases the causal momentum* of the Ss and increases autonomy thereby; this also works for anticipated Ss or what utilitarian behaviorists and economists call "reward", what philosophers call the forward-looking or pull-model features of motives, opposite hydraulic, push-model, or backward-looking features.

There are exceptions, but these sketches constitute initial elements of a system of distance principles that may be construed as "metacausal distance laws",[53] laws of self-regulation, and agents who know how to manipulate them can increase their autonomy.

---

[53] Ainslie (2001) discusses "transcendance" (p. 80) and other distance principles (about how the distance in time between expected reward and its reinforcing behaviors diminishes its effect), and some explicitly spatial distance principles. Behavioral modification certainly incorporates distance hierarchies in its manipulative techniques, altering the distance between S and R as a way of controlling it (Olton and Noonberg, 1980).

There is behavioral evidence for the correlation of hierarchical distancing with what on my analysis is a Humean freedom. Consider counterconditioning, used to cure phobias.

> First the person is trained to relax. Then stimuli which originally elicited anxiety are paired with this relaxation. A critical step in counterconditioning is the use of a *stimulus hierarchy*.... The anxiety associated with a phobia is usually so great that it will completely overwhelm any state of relaxation, at least initially. Thus the therapist establishes a hierarchy of stimuli, all of which are similar to the *target stimulus* (the one which evokes the greatest anxiety) but sufficiently different from it that they produce less of a response. One means of establishing a hierarchy is to vary the distance between the person and the target stimulus or the length of time before which the person must interact with the target stimulus. These spatial and temporal gradients form the hierarchy; as the distance/time between the individual and the target stimulus increases, the amount of anxiety decreases. (Olton and Noonberg, 1980, p. 39; emphases in original)

Olton and Noonberg hold distance in time and space causally relevant to establishing a (phobia-based) S hierarchy, and inversely correlate these with anxiety. Generalizations like these from behavioral science would fill in the gaps in the metacausal theory of autonomy. Olton and Noonberg's description of phobia treatment illustrates their medical application.

> After the stimulus hierarchy is established, relaxation is conditioned to stimuli farthest away from the target stimulus. This ... generalizes to all other stimuli in the hierarchy, lowering the ... anxiety produced by them and making it easier ... to relax in the presence of them. Relaxation is then paired with the next stimulus in the hierarchy until the anxiety response to it disappears. This procedure is continued until the person is able to relax in the presence of the target stimulus itself. (Olton and Noonberg, 1980, p. 39)

The test subjects are trained to relax through biofeedback, a mix of conditioning and meditation – two forms of off-line behavior. Biofeedback's increase in voluntary control over autonomic nervous system functions is evidence of off-line effects. That these *off-line* elements *converge* in behavioral science, *particularly with the treatment of behaviors that render agents unfree in a broadly Humean sense*, is predicted by, and so empirically confirms, the metacausal theory.

Platitudes support distance laws. Consider the platitudes that one should take a deep

breath to calm down, count to 10 before reacting, or "sleep on it" before a major decision. The mere distance in time creates enough psychological space between I/O for Agent to escape spur-of-the-moment reactions that may be cause for later regret. Any psychological mechanism that operates off-line contributes to the liberating effects of causal distancing that attend their causally removed, 2nd-order functioning. Time lag platitudes capitalize on the fact that a temporal gap must obtain between the point at which any mechanism goes off- and then comes back on-line. Taking extra time – temporal distance – enables one to run the troublesome decision through more simulations, thus making the R more mediated, more likely rational, etc.

One more thought on distance is in order, about neural distance. Walter (2002) discusses the role of the prefrontal cortex in simulation, but also in the mental Darwinist theories of Changeux and Dehaene (1989, 1995). According to Walter,

> They distinguish three types of neuronal representations: (1) percepts, (2) images, concepts, and intentions, and (3) prerepresentations. Percepts consist of a correlated activity of neurons that is determined by the outer world and disintegrates as soon as external stimulation terminates. Images, concepts, and intentions are actualized objects of memory, which result from activating a stable memory trace.... Prerepresentations are multiple, spontaneously arising unstable and transient activity patterns that can be selected or eliminated. (2002, p. 569)

What is interesting about this is that neuronal percepts are *directly exogenously determined*, whereas the other two neuronal representations are at least *endogenously mediated*: Intentional objects require neurosignatures in memory, and pre-representations are the sort of potential-pattern elements that go into linking the other two through a process of "resonance" (*id.*), analogous in our brains to what Dennett (2003) refers to as evolutionary "free-floating rationales", like "memes", the cultural analogues of genes.

More important is that our metacausal analysis may be extended *beneath* the purely mental to the neuronal level. That is, the neuronal representations that are directly

exogenously caused – 'percepts' – are close to the bottom of our causal/cognitive hierarchy, way below what we have been calling 1st-order desires or 1st-order mental states. Two speculations arise here: Perhaps the free-floating 'prerepresentational' states are even more rudimentary, though endogenous; perhaps the percepts are what Kant would say are sensory contents caused by outer sense receptors, and the prerepresentational neuronal states are what he would say are their inner forms or category repositories. We may call the former "1st-order psychoneural states", and the latter "2nd-order psychoneural states", for the former are exogenously determined, and the latter endogenously determined, and ours is a causal hierarchy. Both have what Fodor (1992) and Searle (1992) have called "aspectual shape", meaning they neurally instantiate the sort of mental state elements or building blocks that go into things like qualia or intentionality.

These rudiments of mentality are psychoneural bridges between mind and brain. Other neural states, by contrast, do not subserve mentality, i.e., they have no aspectual shape and do not contain any representational, sensory, or other phenomenological content, but are involved in lower-order biological functions, say, respiration, digestion, etc. The third kind of neuronal representations directly instantiate mental states, "images, concepts, and intentions", and so may be called "3rd-order psychoneural states" or "1st-order mental states". 1st-order desires, then, are either 3rd- or 4th-order psychoneural states, and metadesires are 4th- or 5th-order ones, etc.

The conclusions I have been setting up may now be made clearly. The *distance* between the on-line exogenous environmental determination of lower-order psychoneural states and the off-line endogenous determination of higher-order mental states admits of a *causal hierarchical, metacausal* analysis of the sort I have described. That distance is *greater* than previous descriptions suggested. The architecture for this is evolved off-line

processes. *Off-line processes break the causal/functional chain of exogenous psychoneural determination of mentality and volition, loop functions endogenously within reflexive centers of consciousness and will, and enable organisms to access psychoneural free-floating rationales and long-term-memory-encoded preferences.* Simulational processes would augment these abilities in intuitive ways.

These are concrete neurophilosophical claims with straightforwardly-testable empirical consequences, the applicability of which to a theory of autonomy is obvious. Self-regulating organisms are organisms with off-line psychoneural and higher-order meta-abilities. These abilities are natural, observable, quantifiable, and amenable to theoretically-minimalistic causal/functional/behaviorist analyses, and their psychoneural bases are supported by the latest evolutionary theories in brain/behavioral science. This is our major addition to Dennett's "just-so" story of autonomy and our major causal/functional addition to Frankfurt's model.

This is a concrete metacausal self-regulation/autonomy power we possess and animals lack. We can *form* metadesires by simulating belief/desire-inputs, seeing which action-outputs they yield, and evaluating them. We can also *form* Frankfurt's (1971) "2nd-order volitions" – *effective* 2nd-order desires – by running the metadesires through deliberation-oriented 2nd-order simulations, which may assign some metadesires a higher motivational value as a result of their being favored outputs of such simulations.[54] An alternative, adapted from Watson (1975), would be to assign evaluative weights to the output or products of off-line deliberation on 2nd-order desires. This model provides for a hierarchy

---

[54] Our greater imaginative powers are arguably simulational, and arguably make not only prudence, but altruism, possible. Nagel (1970) claims that the key to altruism is our ability to see others as equivalent with ourselves *as persons*; simulation makes this equivalence first-personally accessible and thus psychologically real (Goldman, 1993).

with *different causal powers at each level.*[55] With the distancing between levels that fits with platitudes about self-control, it makes not only agency but a conflicted hierarchy possible, and thus supports Frankfurt's distinction (1971) between problematically unfree agent and wantons.

While there is a link between freedom and metamentality, metamentality doesn't guarantee freedom. For compulsions may rise from lower- to higher-order states. Frankfurt's model permits 2nd-order junkies (Watson 1975). Folks like Hamlet or Woody Allen seem highly reflective but equally neurotic. If metamental states can be dreamed, then freedom doesn't *always* result from metamentality. But the off-line theory does not claim that greater awareness is *sufficient* for greater autonomy, but only that greater awareness of the causes of one's actions is *necessary* for greater autonomy.

The general form of the theory is more important than these auxiliary principles; it is enough to lay out the theory without them. I defend *specific metacausal principles* in the next two chapters, PAPW and PAPM, that rule out counterexamples like "dreamed, thus ineffective metacausality" by restricting the idea *that metacausality is necessary and sufficient for freedom*, to the narrow claim *that metacausality is necessary for freedom.*

Deliberation involves a combination of off-line processes, and its reflective character places it at a temporal distance between its I/Os. It also may involve simulations or off-line functions of emotive, evaluative, and other states, of ourselves or others, and of others' simulations of ourselves. These may arise under simulated I/Os falling under any of these categories, whose outputs have weighted values – attached higher-order pro-/con-attitudes.

---

[55] Rosenthal (2000) makes the claim that it is *uncontroversial* that the causal powers of mental states must be a function of their contents. If so, and this seems right, then *metacognitive states must have metacausal powers by virtue of their metastructural content.*

This manifold of off-line functioning constitutes a meta-structure that accounts for images of freedom we associate with the pre-philosophical phenomenology of deliberation. Consider the image of *the rational thinker*, removed from the causal stream, engaging in counterfactual and other imaginative projections, weighing options, RFAs, and probabilities without being pulled or pushed by internal or external forces. Our model accounts for this and other inflated, phenomenologically-based, pre-philosophical or folk impressions.

Reactive attitudes, like regret, may also be explained on this model. I can simulate others in circumstances similar to mine to see if they would feel regret, by adding and subtracting conditions peculiar to me, such as that I attended Catholic grammar school. Related to this, the reactive attitudes may be rationally reconstructed and revised in a similar fashion. Take cross-level half-heartedness, e.g., desires at one level, repression at another, nonidentification with the repression at another. Stump's lapsed Baptist (1993) may have a 1st-order desire for an occasional social drink, but a 2nd-order aversion based on religious upbringing, and a 3rd-order wish that she could overcome the resilient 2nd-order aversion, given the weakening of her religious beliefs. Thus, if she has a drink and feels shame, she can simulate others who have or lack the intentions she has at each level, and compare whether or not they result in equal shame, and use these simulations to rationally reconstruct her reactive attitudes towards herself (whether she ought to have shame). Similar moves may be used for the other reactive attitudes.

Awareness of the bases and natures of our beliefs, desires, values, characters, and choices – their formation, the processes leading to identification and alienation, their plasticity, etc. – is clearly a first step toward their control or modification, just as in biofeedback awareness of the heart rate is a first step toward its control, however indirect the control over such an autonomic nervous system function that may be. *Reflexive feedback*

systems don't guarantee it, but make control possible. They *may have* evolved from primate simulation mechanisms that made both metamentality and off-line states possible.

We have reflexive, self-altering abilities; animals don't. Some of us are more aware of the control-levers in metamentality than others, are more autonomous. Autonomy is a matter of degree of control. Some have great control over appetite, but not emotions, or vice versa. It is determined that some are more self-determined than others. So?

That off-line behaviors are not all *equally off-line* is irrelevant, for 'off-line' is *relative.* Importantly, our causal powers are concrete, and they are sufficiently off-line *relative to on-line animal causation* for the purposes of the off-line theory. There are degrees between involuntary, voluntary, and maximally self-regulating behavior. Most animal behavior is involuntary, though the voluntary/involuntary distinction is relative. Most people, by stark contrast, usually act minimally voluntarily, and have a degree of autonomy sufficient for responsibility. This holds though most have gone through habituation processes that form stabilized character, thereby limiting the range of options found salient or desirable. But most can alter behavior enough to be responsible for who they are and what they do. Some are more or less free.

These judgments are highly contextual. In the next two chapters we formulate a theory of the will and of free will designed to explicate our divergent intuitions in the broad range of cases that constitute the data to be explained by any theory of the will or of freedom.

Chapter Four: The Metacausal Will

Let me first apply the causal/functional cognitive-architectural model more closely to the issues in the free will debate, then to the construction of a theory of will itself. Our task was to identify the abilities that comprise autonomy, how we have them, and why some think we don't. To tell a "just so" story of how they could have arisen naturally is to show that they are compatible with determinism. Dennett's just-so story traces complex phenomena like thought to tropisms, and identifies language, reflexivity, and metacognitive phenomena as crucial in the development of powers distinctive of autonomous persons. My addition to his just-story is the suggestion that our off-line/metamental abilities may have evolved from primate simulation, but I also add metacausal analyses of these same powers, regardless of how they may have arisen. I also add the claim that autonomous agents have metacausal control over their actions.

Like Searle (1992), I view mind as a higher-order neural state. This easy-approach to mind sidesteps the mind-body problem in one move (1992). G. Strawson has also argued that if materialism is true, it follows that the mental is physical (1994, 1997). The plausibility of these views further justifies my rejection of the notion of the impotence of the mental. I will argue that autonomous or *narrow* metacausation is top-down, from mind to body, and autonomy is a metacausal power of the metamental features of the brain that enable agents to significantly author their actions. Agent is free *iff* he acts narrowly metacausally. A sketch of this type of autonomy may be garnered from Levin's adequacy conditions on an account of freedom:

> We need adequacy conditions, neutral with respect to all competing analyses, on proposed explications of [freedom]. [Freedom] is whatever, if anything, satisfies these conditions, and the claim that we *are* free is the claim that something about us does satisfy these conditions. I suggest the following two claims as just such a neutral statement of what we are talking about when we say 'man has free will':

(1) Man's causal influence on the world differs from that of other things.
(2) We are sometimes in control of what happens to us. (1979, p. 237)

Metacausal control satisfies these conditions. Animals, young children, and dysfunctional adults lack this ability, so it is grounds for excluding them from the class of free agents, and for including ordinary adults. While competing criteria are open to counterexamples, autonomously-functioning adults do not suffer from, say, indoctrinated religious repression of healthy sexual desires.[56] Competing criteria are open to a regress problem, but since the control identified by our theory is top-down from Agent's own conscious will, the very notion of a downward-causation regress from higher levels is incoherent. Thus, *narrow metacausal compulsion* is oxymoronic. Other forms of compatibilism, e.g., the reasons-sensitivity account,[57] appear incomplete.

Free actions standardly involve consciousness, construed by Rosenthal as a thought *about* a thought – a higher-order thought ("HOT") (1991b, 1997, 2000). The HOT theory holds conscious states meta*intentional*, as opposed to *metasensory*, because there is no sense organ for discerning our own mental states, and the only alternative is that we are conscious of them by having thoughts about them (Rosenthal, 2000).[58] This dichotomy is false, for there is no need for eyes *in a dream*; sensory states need sense organs only in their etiology. Conscious "sensing" need not require sense organs to be nonconceptual.

I agree with Rosenthal that consciousness is *metamental*. *Any/all* metadesires or

---

[56] To be fully metacausally functioning is to not be cognitively/conatively "stuck", which is why metacausality may subsume reasons-sensitivity as a species of autonomy, and not the other way around. On the relevance of being "stuck", see Sober (1995), pp. 326-335.

[57] E.g., Fischer (1994); F&R (1998); Wolf (1990). Shatz (1986) attributes "reasons-sensitive" to Levin (1979), but Levin doesn't reject autonomy (1979, 2003); F&R and Wolf do.

[58] *Cf.* Güzeldere (1997). "Intention" may involve one of two distinct senses: Brentano's *aboutness*, or teleology (goal-orientation). All actions involve an aboutness, but not vice versa.

metathoughts are "metaintentions". But consciousness need not be meta*intentional*, for either mental state level may be purely qualitative; e.g., *a purely qualitative awareness of the sensation of red.* But all metamental states are meta*cognitive*, for they are cognitive states that are *of or about* sensory, conative, or emotional states. So, properties attributable to metacognition need not be attributable to metaintentionality.

The mental functions that matter are causal – not, as Frankfurt holds, motivational. Syntax alone differentiates between the *meta*causation in *meta*cognition and the *causation* in *cognition*, since *the causal powers of mental states are a function of their content* (Rosenthal, 2000), and only the content of the former are tiered. The metacausal analysis, however, is independent of a particular analysis of causation. Similarly,

> the kind of position I want to advocate here concerning macrocausation is largely independent of the particular views concerning the analysis of causation. Moreover, I do not wish to tie the fate of my general views about macrocausation too closely to the fate of my proposal regarding a proper construal of the relation between macroproperties and the microproperties on which they "depend". (Kim, 1984, 263)

But, nevertheless, I simply *grant* the pessimist's strongest version of determinism, *viz*, the Laplacean/Newtonian form, the nomological model assumed in CON and amenable to the Hempelian covering law model. My strategy is to show that determinism, even in its most rugged cloth, is autonomy-friendly. Chapter 2 rendered this compatibility plausible.

Hobbes thought the voluntary was motion in the imagination; metacausality is causal motion in metamentality, either laterally or downward, as when the metadesire to have the desire to love the law causes the desire to love the law. Metacausality thus permits determinism-friendly 'agent-causation' as top-down metacausation. Hierarchical accounts are considered compatibilist, even by their critics.

> For the most part, advocates of hierarchical accounts see themselves as continuators or rehabilitators of classical compatibilism. (Shatz, 1986, p. 451)

One reason is that there is nothing inherently indeterministic about metastructural properties.

Hume held that unconstrained desires are caused but free; hierarchical accounts attempt to explain why/how they are free. Hierarchical features that might be the basis of a caused free will are (1) accord between levels, (2) accord plus some special relation between levels, or (3) absence of inter-level discord (Shatz, 1986, p. 453). The unique feature in my account is higher-order *causal control* over lower-order causal elements. Since my hierarchical account is intrinsically causal, it is compatible with causation. Since Hume advances an account in which some causes are free, and since my account is intrinsically causal, it is neo-Humean.[59]

If freedom is a species of causation, it is empirical, not speculative. Some think nothing empirical can bear on freedom, and laugh at theorists like Libet (2002) who think otherwise, as if the idea is on par with empirical tests for Cartesian mind or *elan vital*, say, weighing a person before and after death. Some hold a Kantian anti-metaphysical view about freedom, as a paralogism, or for other reasons. But, as Hume showed, freedom may be treated empirically by referring to *causal* differences between folks with/out compulsions. His model is proto-empirical, but empirical work in metacognition is relevant here.[60]

A 'geologic *column*' illustrating epochs in geologic time is not an *actual column* cut into the geological layers to be found stacked together in the Earth. Rather, it is an *interpretive construct* informed by data from astrophysics to chemistry. Likewise, a 'causal *column*' illustrating stages in meta/causal history is not an *actual column* cut into the layers

---

[59] Most accounts focus on features of intention relative to which causation is extrinsic. Katz's 'new intensionalism' focused on *intrinsically intensional* features of terms, claiming most forms of intensionalism focus on extensional (extrinsic) features (1997). Soft *determinism* ought, by analogy, to be intrinsically *causal*; metacausality is intrinsically causal.

[60] See Rosenthal (1997, 2000) for a review of empirical research on metacognition.

of causation to be found stacked together in, say, the mind/brain, or inscribed into its historical record in the metaphorical 'rings of the trunk' of the evolutionary 'tree of life'. Rather, it is an *interpretive construct* informed by data from evolutionary biology to conditioning theory. This column reflects causal strata: At one end are purely mechanical types, such as that involved in Newton's billiard balls; at the other end are free actions; and in between are life,[61] homeostatis, locomotion, brains, mental states, and metamental states.[62] Empirically significant co-varying properties map onto the causal column; informative generalizations result. Though topically scattered, many empirical studies cited herein support the idea that metacausation is central to freedom.

Chapter 3 focused on the *functional* features of off-line states, seen in simulation and other phenomena. Their *causal* properties are brought out by analysis of *metamentality*, as in Frankfurt's metadesires and Rosenthal's HOTs. Metacausation is a counterfactual causal power to engage off-line functioning, a power that need not manifest as action in every case. Not all off-line activity is metacausal. Dreams are off-line, but not necessarily metacausal.

> By destroying neurons in the brain that inhibit movement during sleep, researchers found that sleeping cats rose up and attacked or were startled by invisible objects – ostensibly images from dreams. (Winson, 1997, p. 59)[63]

This *on-line response to off-line stimuli* suggests a neural switching system that toggles on-/off-line functions. Research suggests dreams serve an economizing processing purpose.

---

[61] Paul Taylor (1986) marks *life* off as unique in nature, and refers to the novel, unique axiological property of living things by calling them "teleological centers of life".

[62] The points of significance that will turn up on the causal column will be similar to those that appear on the emergentist's history of the world, leading from lifeless matter to consciousness. See Langer (1967); *cf.* Humphreys (1997). As Kim points out, however, a deflated version of emergent, namely *resultant*, poses no problems for physicalism (1989).

[63] Hobbes' hydraulic model of motions in the imagination applies to dreams, a function of "diminishing" sensory motions in through the system. Their diminished power is a function of spatiotemporal distance. McGinn (2003) endorses a hydraulic, Hobbesean view of dreams.

> For higher capabilities to develop, the prefrontal cortex would have to become increasing large – beyond the capacity of the skull – unless another brain mechanism evolved.
>
> REM sleep could have provided this new mechanism, allowing memory processes to occur '*off-line*'. Coincident with the apparent development of REM sleep in marsupial and placental mammals was a remarkable neuroanatomical change: the prefrontal cortex was dramatically reduced in size. (Winson, 1997, pp. 63-4; emphasis mine)

Hallucinations, fiction, memory, and language-dependent thought involve off-line elements. Otherwise identical 1st-order mental contents/states with different 2nd-order attitudes about their on-/off-line 1st-order status generate *radically different causal powers*. Seeing a hallucinatory state as on-line – 'real' – leads to psychosis; seeing it as off-line can be empowering, as when a schizophrenic realizes a hallucination *as such*.[64]

Though not all off-line processes are metacausal, *narrow metacausal power is power to control the levers that place lower-order mental/causal sequences on-/off-line*. Decision and its executive power to initiate action are functions of the ability to go on-line after deliberation, and restraint is the ability to stop an on-line process from functioning, to take it off-line. 'Off-line' and 'hierarchy' are contrastive, relative to the on-line processes and n-tuples upon which their meanings are parasitic, respectively. Thus, some off-line functions and hierarchies will more transparently instantiate those concepts than others, particularly *more isolated* off-line functions and *more stacked* hierarchies. Though we began our just-so story with simulation as illustrative, these abilities may have arisen through any variety of evolutionary processes, such as language, memory, or dreams, each of which has enabled some ability to engender a metacause.

Compare the causal differences between the wills of animals and people. An

---

[64] Lower-order animal states lack judgement – a deficit that explains those who "freak out" on LSD as lacking the higher-order *cognitive* ability to view the hallucination *as off-line*.

animal's will is nonhierarchical: 1st-order intentions are *directly* determined by genetic/environmental factors. An animal is genetically predisposed to aversions, attractions, etc., and congruent objects/events directly engage its intentions in immediate S/R sequences. Since the direction of the transference of motion is from environmental S to mental/behavioral R, it is bottom-up causation. The world/mind contrast need not be dichotomous for the *bottom-up* notion to stand, for the determinative locus of such behavior is nonmental, exogenous. There is no psychological space between an S/R pair for an animal to ponder. The strongest motivational potential simply issues immediately in behavior. Such a will is fully on-line with the environmental causal nexus. Animal causation is lower-order, exogenous, but agent causation is higher-order, endogenous.

> Agent causation is a rather mysterious doctrine.... Can we find a naturalistic account of the self ... that avoids this "obscure and panicky metaphysics" while distinguishing agents sufficiently within the causal fabric? (Dennett, 1984, p. 76, quoting Strawson (1962))

The metacausal theory does that, so it avers a form of *determinism-friendly* agent causation.

Dennett says an attribute is natural if evolution *could have* produced it; this is a "just-so" story. Dennett gives just-so stories for language, consciousness, and other meta-abilities (1984; *cf.* Humphrey, 1982). Since metacausation arises with these abilities, he indirectly showed it is natural. Folk ascriptions that involve mental states (the ascriber's) about mental states (the target's) involve *some* off-line metamental functioning. In Dennett's just-so story, the emergence of thought dovetails with that of language as a kind of talking to oneself *sotto voce* in which the brain's formerly-partitioned speaking/hearing functions are linked for the first time (1984, pp. 38-43); he says consciousness may have arisen thus, in tandem (1984, p.

42, n. 25).[65] Thought/consciousness would be intrapersonal communication with dual (sender-receiver) functions on single representations, hence metacognition. Apart from psycholinguistic history, most animal mental states are *of or about* objects/events in their immediate environments. Once an animal acquires a language, however, it is empowered to have mental states about objects/events not immediately perceived.[66]

Searle (1994) argues that language-lacking animals need *only* lack *language-dependent* mental states. Call mental states that do require language "linguistic states"; examples include those *about language* (e.g., translation), linguistically constituted (e.g., "I have $20"), whose complexity requires language (e.g., thermodynamics), that link the thought to an arbitrary system (e.g., April 30, 1993), or that represent facts so remote as to require language (e.g., that Napoleon ate good food) (1994, p. 213). Language permits states not about the vicinity, e.g., the thought *that Superman is from Krypton*. Pains, vision, and fear do not require language, so we may ascribe these to animals. States immediately dependent or not on the environment are, respectively, on-/off-line. All language involves metacognition, a *stand-in* associated with the state in which one is conscious of its referent. This supports the causal column claim that higher mammals have some metacognitive powers, and that humans have more. It is no surprise we are autonomous, for within the off-line space of language, thought, and consciousness we find the final specs for elbow room.

---

[65] This is inconsistent with his denial of the 'Cartesian Theater', the place in mind in which mental experiences occur. But there can be no *unmediated communication*, so the medium for intrapersonal communication can be a Cartesian Theater, one that makes elbow room built into the very structure of thought (which, for Dennett, is lateral – mental/mental – mental causation).

[66] There are alternatives, e.g., memory. Nonbaptismal language requires memory – *re*cognition of the meaning relation. *Re*cognition of any sort, arguably, involves a metacognitive element. All our concepts rest on *re-cognition*, thus on patterns in experience; see Price (1953). If so, *meaningful* experience involves metacognition – identifying *this* experience as the same as *that* one (true in all linguistic mentation), however preconscious. As our ability to store 10,000 words in computers facilitated complexity of thought, so also did the appearance of language.

Endogenous mentality/action is built into the metastructure of metamental causation. In downward causation, intentions result in neuromuscular activities.

Consider LeDoux's research: The *fear* of seeing a snake is processed faster than the sight of it, which traverses a longer path. LeDoux calls the former "emotional memory", the latter "declarative memory" (1997, pp. 74-5). Emotional memory is optimally unconscious, but "[d]eclarative memory involves explicit, consciously accessible information" (p. 72).

> Pathways originating in the sensory thalamus provide only a crude perception of the external world, but because they involve only one neural link, they are quite fast. In contrast, pathways from the cortex offer detailed and accurate representations, allowing us to recognize an object by sight or sound. But these pathways, which run from the thalamus to the sensory cortex to the amygdala, involve several neural links. And each link in the chain *adds time*. (1997, p. 74; emphasis added)

The more *environmentally* significant, the less mediated – the less *temporal distance* in – the neural processing. *Meta*cognition is more *metamentally looped*, so more *neurally mediated* than cognition. More mediation means more *temporal distance*, *more time* for reflexive choice.[67] Metamentality loops neural functions, so it should be traceable in off-line, *distance-rich* looped neural networks that realize self-regulative consciousness – the grey matter of elbow room.[68] Consciousness is unnecessary for rapid/unmediated processing of the fight/flight response, yet it empowers one to restrain instinctive in favor of rational responses. Deliberation is rational review of alternatives, weighing short-/long-term goals, the ranking of action/outcome pairs. Even on a Hobbesian model (*cf.* Ainslie, 2001), it involves the ability to transmute animal-type conative propensities and environmental inputs.

Frankfurt states that *whatever* intentions are effective are "volitions" and that

---

[67] Hamlet aside, factors we are *aware* of are brought within the "horizon" (Levin, 1979) of options. Dennett (1984, p. 69, n. 31) says there is a point of diminishing returns in competitive contexts like evolution, which favors limited-range meta-thinking, "heuristic" deliberation.

[68] Crick suggests the seat of autonomy is "in ... a region receiving many inputs from the higher sensory regions and ... at or near the higher levels of the motor system" (1994, pp. 267-8).

freedom involves meta-volitions; he fails to give a substantive account of them. But conflict between intentional orders is not the only way to exercise the will,[69] for anything that exercises the metacognitive mind may exercise the metacausal will. Meta-volition is not the only route to autonomy; metacausality is more all-inclusive, and sheds light on personhood.

The structure of the will is embodied in idiolectical *psychoneural* pathways that instantiate the labyrinthine configuration of the metacausal mind.[70] Autonomy links to identity because persons are individuated by conative identifications/alienations (Frankfurt, 1987), and their value rankings (Watson, 1975) are soft-wired into the metacausal structure of their will, and its associated wet-wired neural underpinnings (Benson, 1996). As Frankfurt (1987) put it,

> [T]hese acts of ordering and of rejection – integration and separation [of approved and disapproved motivations –] create a self.... They define the intrapsychic constraints and boundaries with respect to which a person's autonomy may be threatened.... (p. 170-1)

The self-creation of approving of one's desires links freedom, identity, and value.

> The normative component of reason-sensitivity gives some substance to the metaphorical identification of a man with his values. A man embodies his conception of what he *ought* to be to the extent [that] his desires are his because they have passed the test of his reason. He is, to that extent, a self-created self. (Levin, 1979, p. 249)

Frankfurt further develops this link between the hierarchical will and personhood:

> [A] function of decision is to integrate the person both dynamically and statically. Dynamically insofar as it provides ... for coherence and unity of purpose over time; statically insofar as it establishes ... a *reflexive or hierarchical structure by which the person's identity may be in part constituted.* (Frankfurt, 1987, p. 175; italics added.)

---

[69] Though inter-level *conflict* is *vivid*, the will is conflict-free and thus *less vivid in its ubiquitous appearance*; *cf.* the *vivid* sense of *disequilibrium* with the mostly *unnoticed* (ubiquitous) sense of equilibrium. Moral-conflict-oriented theorists like Kant, Campbell (1967) and Kane (2002b) ignore its daily (heterogenous) functional abilities.

[70] I highlight "psychoneural" to dispel the implicit dualism that would separate mind and brain. See Searle (1992) and Northoff (1999) for a defense of mind-brain monism.

Our deepest values/ends – static normative/motivational structures – are realized via dynamic decision-making in the psychoneural configuration of reflexive mind. Metacausality is central, but only *top-down* metacausality is autonomous; this is *strong/narrow* metacausality.

Like Goebbels or Pharaoh (Stump, 1993), suppose fanatic "Q" wants to harden his own heart against uncharacteristic compassion for his enemies. If indoctrination is internalized and Q uses it to sculpt himself as such, he displays the sort of self-regulation my account identifies as autonomous. But Q seems intuitively unfree; his beliefs are *overdetermined* by indoctrination with which he identifies. The genesis of the indoctrination is key: Did Q elect it as a young idealist as prescriptivists say we choose ultimate values, or was it spoon-fed at a pre-agential stage? Was Q brainwashed, or socialized by community? Causal history matters – what F&R call 'tracing' (1998). How much uncoerced sculpting came from Q's pre/agential "self-forming" stages matters to compatibilists (Mele, 1995) and incompatibilists (Kane, 2002b) alike.

If another agent controls Q, Q lacks *agential* self-control. For even if *some* self-control is present, if it is surreptitiously or insidiously manipulated by another agent, it is then not *Q's agency* that is responsible for his behavior. If there is no agency behind Q, but bad luck, then Q is just unfortunate. If Q was brainwashed, changed, and basically restructured as an agent, he loses some responsibility for self, but maintains responsibility as an agent with minimally-sufficient metacausal powers. Metacausal beings author some of their mental states,[71] and so sculpt themselves. They consciously hold the reins to their modes of being/acting; they are autotelic. Wantons' modes of being/behaving are fully

---

[71] Non-voluntarism is tautological for purely alethic beliefs, whose content is not constituted by the believer's mental states. But not all beliefs are purely alethic, so some may be voluntary. But we can *indirectly* alter even alethic beliefs; see, e.g., Elster (1979).

determined by lower-order mental states, and these are fully determined by genetic/environmental conditions.

It is productive to focus on *metacausal* states; once we isolate their functional properties, we may look for neural links with activities such as simulation, as foreshadowed in Locke (1690):

> For, the mind having ... a power to *suspend* the execution and satisfaction of any of its desires ... is at liberty to consider the objects of them, examine them on all sides, and weigh them with others. In this lies the liberty man has.... (*Essay*, II, XXI, p. 48)

The power to *suspend or execute* is to go *off- or on-line*; it covers most cognitive processes. Akrasia is noncontrol over conations, failure to keep sub-optimal conations off-line.[72]

Skilled meditators illustrate control over autonomic nervous system functions and can uncouple cognitive/somatic processes. Trance-invoking techniques inhibit somatosensory responses to environmental stimuli (e.g., proprioception), which "neurophysiological habituation" explains the phenomenology of transcendent experiences (Anand *et al.*, 1961). This top-down ability disengages mind from lower-order, S/R-type causation. Nozick's soundless-stereo example describes meditation as "an unusual procedure of experiencing wherein most but not all functions are damped down" (1981, p. 159). Meditation thus involves the ability to switch cognitive/conative functions on/off, to go on-/off-line.

If the neuronal body-matrix theory designed to account for phantom limbs is correct, then perhaps meditators just inhibit this neuro-matrix. In discussing phantom limbs, Melzack states,

> In essence, I postulate that the brain contains a *neuromatrix*, or network of neurons, that, in addition to responding to sensory stimulation, continuously generates a

---

[72] Recent popular studies link addictions and kleptomania – paradigms of unfreedom – to serotonin re-uptake failure and related neurotransmitter problems. This suggests inability to switch between on-/off-line functioning is due to a *blown neural fuse*. Analogy: The driver hits the brakes, but the car doesn't stop, a psychoneural, top-down mental causation linkage problem.

> characteristic patterns of impulses indicating that the body is intact and unequivocally one's own. I have called this pattern a *neurosignature*. If such a matrix operated in the absence of sensory inputs from the periphery of the body, it would create the impression of having a limb even after that limb has been removed. (1997, p. 87; emphases added)

The brain seems to map perceptions onto a body-matrix neural construct that explains phantom limbs and also yogis' extra-corporeal experiences, confirmed by the homologous brainwaves of meditators and sleepers. The yogi takes the body-construct off-line, as the sleeper's brain does.

Of dreams, Winson (1997) asserts that the ability to take somatosensory and other cognitive/conative processes *off-line* serves a simple survival purpose: to inhibit movement in dreaming animals. On his analysis of the somatosensory dissociative process that occurs during dreams, the brain in REM/dream sleep inhibits motor neurons associated with locomotion and proprioceptive mechanisms normally operative in the feedback-loop for orientation, or else the organism would engage in locomotion while unconscious in response to dream content – as occasionally happens when a dog is seen kicking in its sleep.

Recall Winson's remarks (above) about the evolutionary purpose of dreaming in allowing memory processing to occur '*off-line*' and the reduced prefrontal cortex it requires (1997). Greater attentional ability is correlated with lesser neural activity, ability to *narrow down* cognitive functions to essentials (Hamilton, 1976, 1981; Hamilton *et al.*, 1977, 1984); this is top-down, metacognitive-to-neural causation. Epileptics who use biofeedback to produce *any* alteration in brainwave patterns *decrease the frequency/duration of seizures* (a paradigm of constraint) (Olton and Noonberg, 1980). This suggests that *any increase in metacausal control might decrease the frequency/duration of constrained states*, but this is an empirical matter.

Since organisms are *locomotory*, they are off-line *relative to* billiard balls and objects

that are *only* directly and immediately subject to laws of motion, and as such are entirely exogenously determined. The causal difference between entities only subject to laws of motion versus other forms of determination maps onto the difference between living/non-living things. Billiard balls are *globally* on-line and lower-order determined; organisms are not *equally* on-line, not entirely lower-order-determined. The more evolved, the more local the relevant causal nexus, and the less on-line relative to global (Newtonian) determinism. This confirms the causal column and supports my objections against global "L" in CON.

World/mind causation involves *exogenous* stimuli that *immediately* determine an organism's I/O states, whereas mind/mind causation involves *endogenous* stimuli that *mediately* determine an organism's I/O states. The causes of metacausal states are mental; the causes of lower-order states are not. Persons are *non-superficially metaphysically* responsible for actions that they *metacausally* author; they are "deeply" responsible (Wolf, 1990), not just superficially – *merely* causally – responsible.[73] Our account of deeply-causal responsibility can ground an account of moral responsibility and related attributive practices.

When objects stimulate the senses, they cause Walter's 'percepts' (2002); this is exogenous, lower-order, bottom-up mental causation, as in Descartes's perception of wax's sensory qualities. Non-sensory intentional states need not be, and so are at a causal remove from the more-immediately-exogenous determination of sensory states. A non-sensory state might be about wax-substance, not given in perception. Even if intentional states are *genetically* related to sensory states, they are still at a causal remove from sensory states.

This causal remove warrants a hierarchical construal of the distinct kinds of mental causation involved (sensory/intentional), though these need not be dichotomous. Sensory

---

[73] Davidson's sort of deviant linkages between reasons and causes (1968) involve superficial causal responsibility; nondeviant links involve deep causal responsibility.

states are caused *almost* entirely/directly exogenously; intentional states are not. So, beliefs with *purely* mathematical, logical, or tautological contents (e.g., 2+2=4, Pv~P, or A=A) have *fewer* immediately environmental factors, in their etiology,[74] than perceptual beliefs.

The less environmental the etiology – the more *distance* between environmental causes/agent effects – the more metacausal, and vice versa.[75] Distance entails that a state's off-line status entails greater neural processing distance from environmental causation. LeDoux's (1997, pp. 74-5) studies support this. If intentional states are interdependent,[76] even lower-order ones are more removed than purely sensory states. Even the non-cognitive higher-order neural processes that subserve mentality, e.g., that go into *binding*, involve greater systemic distance if inherently holistic. If each intention functions systemically, none is *directly* bottom-up determined, but even if not holistic, metacognition is integrated in indexical apperception.

Propositional *attitudes* involve an extra functional step, for there is surely more elbow room in attitude-Ap than with p – whether the attitude is *de re* or *de dicto*.[77] The more elbow room from the greater neuro-complexity built into metacognitive states, the more metacausality. Rosenthal's HOTs-model puts conscious states' higher-order components at a

---

[74] This holds on a Quinean/holistic view of the *a priori* as centrality in our conceptual framework, for the more central the beliefs, the further their causal distance from the "contact boundary" (Perls, 1947) with what Quine (1953) calls "the tribunal of experience".

[75] The more basic the science, the lower the causation-type on the causal column; the higher on the column, the more local the relevant causal nexus referred to in correct explanations.

[76] Intentional content is a function of the inferential roles a belief plays in the system; so, it must be partly holistic, as must be the neural functions that subserve it.

[77] All (pro, con, neutral) attitudes toward p are mental states about p; p is a representational state; *ergo*, all attitudes are metamental, whether the object state signified by "p" is construed *de re* or *de dicto*. Animal mentality lacks room for judgement: Content arises automatically, as if the pro-attitude is hard-wired without any mediation between attitude and content. *Animal mentality is attitude-poor*, so elbow-room-poor. The attitudes enhance elbow room.

remove from lower-order ones relative to which they derive hierarchical status. Akratic states are objects of consciousness. Thus, though all higher-order *qua* mental states about others, not all conscious states are strongly metacausal. Some are about perceptual states, such as being aware that one is looking at an object, still a mental state even if the latter component is *de re*. There is little within one's power about being in a state of being aware that one is looking at an object: The awareness seems a brute fact, as does the perceptual object. And that would make it appear that the state is environmentally determined, a kind of bottom-up causation in which environment causes perception and the consciousness of the perception too – *de re* or *dicto*. If so, then there would be metacognition without metacausality, or metacausality without autonomy.

But experience reveals that the mere fact that one is conscious of perceptual states enables one to redirect one's consciousness elsewhere, and that makes consciousness metacausal. Chalmers' bridge law is that consciousness involves the ability to verbalize and act on the object of awareness, and vice versa (1997, p. 36).[78] Thus, it makes sense to hold that *conscious* perceptual states are metacausality-empowering, one important causal step away from perceptual states that *aren't conscious*.[79] So, consciousness of mental states – metacognition – is a metaphysical condition for the ability to alter them, even if all that involves is verbalizing it or looking away. Attention-training can thus enhance autonomy. Ability to avert eyes is ability to alter perceptual states, aided by attention. Ability to stop drifting in daydreams requires that one 'catch oneself unawares'; this contrasts with the lack of attention/control that attends dreaming. Attention is the empowering variable.

---

[78] Pre-linguistic beings are counterexamples to the verbalizing component, but this can be fixed by dropping the biconditional for a one-way conditional (verbal to conscious only).

[79] Searle's connection principle says mental states are accessible to consciousness. If joined with Chalmers' bridge law, the range of autonomous control covers all mentality.

The metacausal view of akrasia is similar to hierarchical views in terms of higher- over lower-order mental states. Where Agent succeeds, the agency is metacausal; in akratic cases, it is not. If there are compulsive metaintentions, they are not metacausal; atypically, bottom-up causation exceeds its usual parameters. One can calmly/knowingly just eat the negatively-valued cake, but this is just what makes it akratic, since it is negatively valued – otherwise it wouldn't be a problem at all.[80] Meta-powers are voluntary, so one may not use them effectively or at all. It is the meta-power to stop the cake-eating sequence that makes it plausible that Agent is "allowing" herself the cake; she can refrain, but doesn't. Without the meta-power to refrain, her responsibility is diminished, as is the child's or dog's, who lack metacausal control altogether.

If the causation is not top-down controllable, it is not narrowly metacausal, autogenous, or autonomous, so there can be no compulsive metacausal hierarchies. Not every metacognitive state is strongly metacausal, but every strongly metacausal state is headed by consciousness, in line with the idea that consciousness is central to freedom. In the conscious, metacausal space of deliberative rationality, 1st-order options are reviewed, selected, and engaged from a higher-order (Lockean) vantage; this is the metacausal model of elbow room.

For Rosenthal, consciousness is a *thought* whose content is another mental state. But consciousness is often intuitively a *qualitative* or *non-intentional* awareness of another mental state, not a "thought". Awareness would be analogous to sensory qualia – not necessarily intentional. Intention is inessential to the conscious side of conscious states, regardless of its object, so it does not require a *that-clause*, e.g., grief about severe pain.

---

[80] It is doubtful that the experience would not itself reveal an element of negative value, e.g., feelings of self-loathing, alternating between hurrying and interrupting the eating, etc.

Consciousness may function as a *field* in which thoughts/sensations are registered, which would explain why the sensory/intentional are not reducible to each other, and have only consciousness in common, or as Searle says, the disposition to consciousness (1992).[81]

If consciousness is not just another level in an *intentional* hierarchy, then this provides an additional functional difference between sensory/intentional hierarchies and the conscious states that top them. If the top element of a metacausal state is conscious, it is *different in kind* from the element linked with the environment, and so at a causal remove.[82] This provides a non-arbitrary reason to stop iterating intentions, for a hierarchy may be topped simply by consciousness.

Consciousness is the control-lever fed by what long-term memory ("LTM") and its neural grooves identify as salient, forming cognitive/conative *gestalts* from sensory experience before slow-moving conscious attention arrives. Agent's dynamic consciousness and static will stops the regress. Kane (2002b) calls this the "self-system", whereby present choices are not unfree, though they appear so, so long as they result from Agent's prior "self-forming actions". This is fine, but Kane is an indeterminist, and a determinism-friendly version will do.

There are nonarbitrary limits to attention and the number of reflexive-loopings that can be held in consciousness. Dennett's arguments in favor of "heuristic" – time-limit-sensitive – deliberation, as opposed to the "perfect" deliberation that would kill Buridan's donkey (1984, p. 69, n. 31), show evolution favors a limited range of meta-thinking. And so

---

[81] G. Strawson (1997) says for all we know consciousness may function as a "field of forces"; *cf.* Kinsbourne (1992). This idea undermines Dennett's arguments against the Cartesian theater, grounds a Kane-style account, and undermines Rosenthal's senses/thoughts dichotomy.

[82] That sensory states are more immediately-determined than intentional ones helps explain the phenomenological intuition of acausality in deliberation, lacking in animal causation.

for our "last stop: consciousness" reply to the regress problem. The metacausal mind is topped by *the* 'Elbow Room',[83] the conscious, functional-control tower for executing decisions and actions.

For Dennett, agent-based control is just tropisms joined over evolutionary time. He is impressed that we can "go meta-" with feedback-loops, but – like Frankfurt – doesn't specify the *causal/functional* features of looped processes as the bases of autonomy. The metacausal model does: *Reflexive causation* leads to *metacausation*, which allows integrated conscious control of the subsystems of agency. This is a causal/functional ability wasps don't enjoy.

Metacausation comports with Hobbes' idea that actions are voluntary *iff* their causes are in the imagination, which is metacognitive. In the *Leviathan*, exogenous motion is transmuted into endogenous motion in the imagination – "diminishing sense". Compare our idea that environmental inputs may be redirected, as in diverting one's gaze. Hydraulic 'force' is "diminished" the more it moves through mind. Thus, consciousness of exogenous forces diminishes their effect. As exogenous causes work their way into the hierarchy, they are "diminished", transmuted by Distance within Agent.[84]

Autonomy involves the metacausal ability to dis/engage lower-order systems, and explains why ability to raise one's arm is seen as a proof of freedom, and why libertarians and others reject the equation of motive with intention. For we can raise our arm

---

[83] Dennett calls the psychological space of visual experience the "Cartesian Theater" and "argues on theoretical grounds that it does not exist" (Crick and Koch, 1997, p. 25). But the Elbow Room may be located in the Cartesian Theater – not to imply a *literal field* onto which experience is projected (like film), but there must be a *functional* equivalent. The brain schematizes a body-matrix onto which it projects proprioception (Winson, 1997); perhaps this is how it maps experiences in a functional Cartesian Theater. If so, brain *constructs* experience – *constructivism*, not reductionism. If so, autonomy has been *constructed* by natural selection as the brain's reflexive-control tower from which conative-control-levers are accessible.

[84] This transmutation from psychophysical micro-level-contact with the perceptual environment to macro-level metacausation *virtually* cancels out micro-level determination the way micro-indeterminacy cancels out at the macro-level. But a *transmutation* of causation at the meta-level is not a *break with* causation.

intentionally but without any motive other than the intention to do so. This grounds the idea that efforts/willings are causal factors in action, and explains why we are impressed by great works, deeply-authored by Agent efforts.

Chalmers' (1997) bridge law supports the link with autonomy: Metacausal consciousness forms a *gestalt* of higher-order self-regulative-control over the lower-order cognitive/conative abilities/systems accessible to consciousness that make up agency.

> [C]onsciousness ... is a matter of having *certain sorts of ability*. To be conscious is, for instance, to see and hear. Whether somebody can see or hear is a matter of whether he can discriminate between certain things, and whether he can discriminate between certain things is something that we can test.... (Kenny, 1972, p. 43; emphasis added.)

Kenny's view of consciousness as an *ability* links consciousness and autonomy, for autonomy is an abstract collection of cognitive/conative abilities, the levers of which are integrated in consciousness. *Akrasia* is more general than just the inter-level *motivational* failure hierarchicalists say it is. For if *continence* is *control* over *all cognitive/conative or sensorimotor* behavior, e.g., micturation, then *incontinence* is the *lack* thereof.[85]

Meta/causal non/control is the appropriate category for *in/continence*, and extends beyond motivational in/continence, and beyond issues of responsibility, to the artistic, intellectual, and the athletic, explaining how they are deeply attributable to us, and why they are also appropriate targets for reactive attitudes. Consciousness is more on the cognitive/input side, autonomy more on the conative/output side, but without a dichotomy, for they are integrated: We have control over what we are conscious of. Frankfurt states,

> A creature engaged in secondary responsiveness is monitoring its own condition; [thus, it is closer to being in a position] to do something about [it]. (1987, p. 163)

---

[85] Compare the incontinence of a disabled person and a dieter with sweet tooth. The former lacks motor control and the latter does not, but the latter has diminished control over the motivational inputs to her motor activity with respect to seeking, grabbing and eating the sweets.

Feedback aids teleological reorienting; consciousness is a feedback-mechanism linked to motor ability, a combination with survival value. These are zoological facts of organismic locomotion (Langer, 1967; Stelnach, 1976). Consciousness and autonomy integrate the I/O interactions of intelligent negotiation – organismic self-regulation – on the evolutionary road of survival. Some issues are so survival-basic as to bypass consciousness (LeDoux, 1997), but most benefit from being looped through consciousness. I/O-integrated conscious autonomy makes for good predators and poor prey.

This undermines epiphenomenalism, for consciousness homeostatically navigates threats to its functional integrity. Twin-Earth zombie replicas are impossible: Determinism/materialism entail that there must be a physical difference between *otherwise identical* twins only one of which has mental – higher-order neural – states.

> [W]orlds that are microphysically identical are one and the same world.... Even those who would reject this universal thesis of microdeterminism might find the following more restrictive thesis plausible: worlds that are microphysically identical are one world from the physical point of view. (Kim, 1984, p. 263)[86]

Rather than debunk consciousness, the zombie-hypothesis begs the question against determinism. The problem is in the hidden dualist presupposition that the mental is not physical. For instance,

> If the pain is to play a causal role in the withdrawal of my hand, it must [make] use of the usual physiological causal path to this bodily event; it looks as though the causal path from the pain to the limb motion must merge with the physiological path at a certain point.... But at what point does the mental causal path from the pain "merge" with the physiological path? If there is such a point, that must be where psychophysical causal action takes place. The trouble, of course, is that it is difficult to conceive the possibility of some *nonphysical* event causally influencing the course of physical processes. (Kim, 1984, p. 266; emphasis added)

---

[86] Physicalism entails minds are higher-order brain states (Searle, 1992; G. Strawson, 1997). Thus, some mind/brain states cause other mind/brain states, and the causing/effected brain states each may be higher/lower. Physicalism thus implies any atom-for-atom replica of mine is conscious, since I am. Thus, there can be no unconscious zombie replicas.

If mental states are higher-order physical ones, and we can make sense of downward causation; metacausation just is downward causal control.

Most of the empirical evidence cited earlier helps to show how downward (psychoneural) causation is possible and so supports the physicalist view of mind.

> [T]he only direct way of explaining why a general supervenience relation holds ... is to appeal to ... appropriate correlations between specific supervenient properties and their subvenient bases. If these specific correlations are themselves explainable, so much the better; but whether or not they are, invoking them would constitute the first necessary step.... These correlations are logically contingent and empirically discovered; though they are not further explainable, they constitute our ground, both evidential and explanatory, of the supervenience of the mental on the physical. (Kim, 1989, p. 159).

Chalmers' bridge law is a way to conceive specific supervenience correlations, and to support the idea that Agent need not be in an *occurrent* metacausal state to be free, e.g., when reading, since consciousness brings with it a *disposition* to control. This is *counterfactually-metacausal* voluntariness, and makes nonreflective mental states *candidates* for responsibility.

Compare K and K'; both do x: If K does x because he wants to, but lacks ability to want otherwise, K's behavior is minimally voluntary – an improvement over Sphex's, but not fully free. If K' is identical except K' has causal/counterfactual control over his wanting to x, K' is more free than K. In Frankfurt's example (1969), agent A cannot (because of mad scientist B) do otherwise, but is responsible because he doesn't want to anyway. But A lacks causal control over his wants, so A is not fully free, even if responsible. Since B will prevent A in the event A intends to use his powers, A is equivalent to K under B's control, unbeknownst to A. But since A *normally* enjoys the causal/counterfactual abilities of K', and none of A's choices other than x are affected, his inability to not x seems innocuous, but if he lacked such control over *all* his choices, it wouldn't seem so innocuous anymore. *A*

background of full agency matters.

A is not fully free, and conditionalism stands. Frankfurt adduced this case against the idea that freedom requires strong possibilities, access to forking futures. His counterexample suggests that absent strong possibilities, A chose x without B's interference freely. But if A *would have* chosen x had he had the ability to choose not x, then he voluntarily, *but only partly freely*, chose x, but is responsible for x. I argue later that the impression of freedom in this case only makes sense because A is not *normally* K-like, but K'-like. Inability to do otherwise is inability to causally control one's doings. Thus, *full* autonomy *still* requires ability to do otherwise, not just to do what one wants when there are no alternatives.

Were an Intervener ready to prevent motor execution of a choice to move one's body, that would illustrate why metacausal control over one's *entire system* – not just the *formation* of one's intentions – is necessary for full autonomy, and why in Frankfurt-cases intuitions linger to the effect that Agent *lacks full autonomy* though he is still responsible. One could satisfy the metacausal criterion in general without satisfying the conditional one in particular if one exhibits full metacausal control over every feature of one's life except one instance of x that Intervener stands by ready to alter, if necessary. Again, if that is the only one, and Agent would have done it with or without Intervener, there is little threat to autonomy, but if most or all actions are so, there is a big threat to autonomy. Given PAP-inspired worries about alternatives under determinism, satisfaction of the weaker metacausal condition would be enough for compatibilism. Put differently, PAPW1 suffices *in a given case* if PAPW2 is satisfied in one's overall system.

Chalmers' bridge law helps explain the relations between voluntarism, full autonomy, and our attributive practices. Premeditated acts merit deeper attributions of *causal* responsibility, thus deeper *normative* ascriptions. The metacausal account is thus an

ontological account, unlike Wolf's *normative* account (1990) on which freedom is a function of reasons-sensitivity. Folks sensitive to and able to act on reasons are free/responsible. Wolf adopts Nozick's (1981) truth-tracking idea – that S's beliefs are epistemically competent *iff* S would believe p if p, and S wouldn't believe p if ~p – and applies it to sensing/acting on reasons. The result is reason-sensitivity as the criterion of agent-competency or responsibility – S's acts are reasons-sensitive *iff* S would act on R if R is the best available reason, and S wouldn't act on R if R isn't. Paradigms of unfreedom, from kidnap victims to addicts, stuck with the reason that moves them, unable to grasp or act on better ones, are not reasons-sensitive.

Wolf (1990) rejects autonomy as either subject to vicious regress or unattainable:

> [F]or [Agent] to be autonomous ... not only must [Agent's] behavior be governable by her self, her self must in turn be governable by [her deeper self] and this must in turn be governable by her (still deeper?) self, *ad infinitum*. If there are forces behind [Agent]..., making [Agent] what she is, then her control of her behavior is [superficial]. But if there are no forces behind [Agent]..., then her identity seems to be arbitrary.... But this would seem to exhaust not just the empirical but the logical possibilities: Either something is behind [Agent], making [Agent] what she is, or nothing is. The idea of an autonomous agent appears to be the idea of a prime mover unmoved whose self can endlessly account for itself and for the behavior it intentionally exhibits.... But this idea seems ... impossible. (p. 14)

Wolf's dichotomy ignores alternatives in which degrees of autonomy and self emerge gradually,[87] as in our discussions about the link between autonomy and personhood as a gradual ego-function of higher- toward lower-order preferences. Autonomy is not unattainable or subject to regress.

---

[87] Her dichotomy ignores gradations in the causal column and idiosyncratic self-stylings that form a metacausal agent. The "self-network" (Kane, 2002) or "self-system" (Dretske, 1988; Velleman, 1992) consists of those characteristics Agent identifies with and alienates himself from in his earlier "self-forming actions" (Kane, 2002). Since there is no *need* for indeterminism here, as actualism has been shown implausible, this self-system suffices on the determinism-friendly Frankfurt-Perls model I have defended. For potential links between the self-system concept and cognitive neuroscience, see, e.g., Walter (2001, ch. 3), Flanagan (1992), A. Damasio (1994), and Damasio, Damasio, and Christen (1996).

Wolf calls "Reason" the ability to track "the True and the Good", adding value-tracking to truth-tracking. Rational compulsion may be desirable in purely alethic contexts, for we want our beliefs to track reality, but not in evaluative or decision-theoretic contexts, for that is where we expect and want some 'play' – elbow room – between reasons/actions. Davidson sees this same space between "theoretical" and "practical" reasoning as what makes *akrasia* possible (1980).

Wolf's rejection of autonomy express a strained Cartesian/Kantian idea. Wolf (1990) avoids this with Dennett's (1984) Luther, so morally good he *must* do good, as an acceptable case of rational compulsion. Acting under Reason is offered as a normative alternative to autonomy. What she adds to the Nozickian, Cartesian, Kantian elements is value-tracking: Responsible/free agents' values/actions are sensitive to and determined by the Good and the True. Value-*tracking* is epistemic, not metaphysical. But since truth-tracking entails alethic non-relativism, it is inconsistent with alethic relativism; *relativistic* value-*tracking* is oxymoronic. Although she attempts to back away from strong ethical cognitivism in the last chapter of her work (1990), Wolf's True-/Good-tracking model remains sufficiently pluralistic, but it remains normatively top-heavy. Self-regulation is not beholden to metaethical theory.

Wolf says autonomy adds nothing to reason-sensitivity; but there are concrete benefits to autonomy seen in, say, attention training. Indeed, reason-sensitivity can be constrained, e.g., an insistence on rationality in arational contexts. Ability to disengage ratiocination is useful; so is ability to simulate reason. One's beliefs/desires may be licit, but 1st-order determined, e.g., K, who cannot operate metacausally. Such angelic beings lack psychological structure. For Frankfurt, 1st-order beings – even reason-sensitives – are wantons, neither free nor unfree. The ability to take any psychophysical mechanism off-line

is a form of autonomy; several such abilities form the psychoneural structure of our autonomy. Martians with twice our sensorimotor systems – e.g., prehensile tails, digitized vocal chords – could double our autonomy, as could species with greater metacognitive abilities. Each is worth having, if for no other reason than in its capacity as an amplifier of our elbow room. As Dennett (1984) says,

> [I] do have a reason for a modicum of 'radical freedom', if it will protect me from the wily advertisers out there who are always trying to 'read my mind'.... [O]ur meta-level control thinking may lead us to want to eschew tactics or control strategies that run the risk of being too fully understood, and hence anticipated, by a competitor.... (p. 66)

Dennett says "poker face unpredictability" has evolutionary merit (1984). Mele (1995) discusses scenarios with reason-sensitives as the easiest marks for a type of 'covert non-constraining controller' ("CNC") who manipulates his prey by manipulating its environment (p. 179). Since autonomy over reason-sensitivity adds to 'poker face' against CNCs and other pattern-studying predators, evolution grounds autonomy, particularly in intelligent social species. Reason-sensitivity is just another autonomy-amplifier.

Since mental states are accessible to consciousness,[88] and can be verbalized/acted upon, they are dispositions to autonomy. This captures the link between freedom and consciousness. Some think autonomy is inconsistent with determinism/physicalism. Two incompatibilist responses to this tension are counterintuitive. The first is *hylephobia*, Dennett's term for fear of materialism (1984, p. 91), and leads philosophers to accounts that flout physicalism, e.g., libertarianism, dualism. The second is *hylephilia*, my term for what Sapolsky calls "physics envy", which leads philosophers to think that if something cannot be

---

[88] This is Searle's connection principle (1992). Fodor proposed a "counter-connection" principle (1992), saying any conscious state could become unconscious, but that violates the unidirectionality of causality in dispositional claims. Chalmers (1996) supports this idea, for any information state, even one with only pain as content, is capable of being acted on.

given a reductive explanation in the vocabulary of physics, then it must be false.[89] A third response is compatibilism; its burden is to show how autonomy arises naturally but supports incompatibilist intuitions. This constitutes an adequacy condition on our error theory.

> It is clear that if such a positive account can be given, it will have to declare the intuitions that support [the] vision of the self as unmoved mover to be a sort of cognitive illusion.... It does *seem* somewhat as if we must be unmoved movers; if it *only* seems to be so, what makes it seem to be so? (Dennett, 1984, p. 76)

My account shows the features of autonomy that *make it seem to be so* – why agents *feel* exempt from the causal/physical realm, free to do otherwise under identical circumstances.

Autonomous agents have a *metacausal* ability that *feels like* an *acausal* version of Locke's "power to *suspend* the execution" (1690, p. 48) of lower-order conations. Metacausality explains these intuitions. Metacauses are higher-order states attended/mediated by consciousness, itself a higher-order state (Searle, 1992) that shifts organismic control to a looped, conscious feature of the brain. This control-shift transmutes lower-order causation at the higher-order conscious level, and that makes it phenomenologically akin to a break with causation, though there is no break. Since we usually do what we want, but aren't conscious of how our wants are determined, we feel like prime movers unmoved. But Ockham's razor favors soft determinism, for one can solve the easy problem without indeterminism.

Mele (2002, 2003) favors a touch of indeterminism in the deliberative process where it would appear rational to want it, in the creative mechanism that generates reasons for deliberative consideration, which might produce a greater range of considerations. If determinism is true, Mele will fall back on his nearly identical soft determinism (*id.*). We

---

[89] Sapolsky says this is the cause of reductionism in special sciences. Studies show excess testosterone causes aggression, minimal amounts are necessary, but in mid-range other factors engender a "permissive effect" (1997, p. 47); so, aggression cannot be reduced to testosterone.

want greater elbow room, and a theory-level reason for wanting this is that if determinism is false, CON collapses. But pragmatic reasons for wanting elbow room are mere Pascalian hopes (*cf.* Honderich's "life-hopes", 1993), and we have defeated CON without indeterminacy. Mele has no *evidence* that the mechanism feeding deliberation is indeterministic, but all that recommends the hope is its desirability *if* indeterminism *might* increase elbow room. Indeterministic accounts have been charged with serving no desirable purpose, so the mere desirability of indeterminism is *relevant*, but does not support the claim that indeterminacy *exists* or is *superior* to pseudo-randomness. The easy problem approach rejects any need for indeterminacy. Dennett shows (1981, 1984, 2003) pseudo-randomness renders genuine randomness unnecessary on all relevant counts.

However, I reject the bluff that quantum indeterminacy does not bear on freedom. For if a single event in world history, other than the first, was metaphysically random, CON fails independently of my arguments, *for then initial conditions and laws do not jointly entail P*. The bluff is that even if micro-indeterminacy holds, it is of such vanishing magnitude as to "cancel out" before reaching the macro level, where psychological determinism still holds. But *any* indeterminacy defeats CON. Even if macro-determinism holds, CON fails, since what grounds the claim that determinism undermines freedom is the idea that decision is caused by the forces of physics. Remove the explanatory power of physics and its nomological closure, and CON falls, as would reductionism absent the closure of the physical domain. Without CON's global-nomological closure, psychological determinism fails, for *there are non-psychological determinants of behavior*. So, too, for other forms of determinism – each is likewise defeated by the others in the absence of the unifying power of physico-nomological closure to integrate them under the all-encompassing L that pushes world-state-instants in fatalistic fashion.

Mele argues for agnostic autonomism (2002), and in this his account approximates my own, but he does not advocate a metacausal theory of autonomy. The metacausal model can work either way, but the easy problem's dialectical simplicity favors determinism. Thus, I *have* a theory-neutral reason for an agnosticism that favors determinism; by contrast, Mele prefers indeterminism, but he *lacks* a sufficiently neutral account that supports his preference. Intuitions about the spontaneity of will are insufficient to support indeterminism.

The heterogeneity of the will, recall, fosters diverse paradigms of freedom: Whim supports an acausal model; deliberation, a reasons-sensitive one. Few address the question of what "the will" consists in *per se*. Frankfurt (1971) distinguishes freedom *of will* from freedom *of action*, and Ainslie (2001) has a comprehensive theory of the will, but little to say about freedom. Frankfurt's model informs mine, but Ainslie's informs my *strategy*, for he lays out what *I take* to be *adequacy conditions* on a theory of will, which I adopt with modifications. In my view, the will is real, something to be *explained* – not explained *away* – by reference to (1) its *substance* (components), (2) its *form* (structure), (3) its *function* (the operations it performs *on* its components), and (4) its *causal powers* (the *causal relationships* possible among these three factors in the production of Agent's behavior). This view of *the will* is resultantism (Kim, 1989) or "pseudo-emergentism", not panicky metaphysics.

Autonomy is a conative/motor-nerve-involving ability of evolved species. Though sensory ability is essential to *orienting/world-tracking* aspects of action, distinctive of conative will is its *doing/world-altering* component. I distinguish species' abilities to *respond*. Watermarks are: tropistic, S/R, teleological, prudential. Chemical features are absent at the atomic level, and at the organic level lies an increase in structural complexity – life. If we map kind-of-responsiveness to kind-of-entity, the same increase attends shifts from mineral to plant, insect, mammal, primate, human. Increased causal/functional powers

and structural complexity co-vary and define the causal column. The *kind* of causal interactions between hydrogen and oxygen – a function of valence-structure – are not the *kind* between a chess player and a computer – a function of algorithm-structure. We may reduce design to mechanism, but the *explanation* is often lost thereby (Schiffer, 1990).

Contrary to macro-level-eliminative reductivists (e.g., Kim 1992, 1999, 2000), piles of rocks have pile-level ("L2") causal powers that their members ("L1") lack; e.g., they are impenetrable by other L2 structures. These L2-powers pseudo-emerge naturally from the *relational arrangements* of L1-constituents, not from their *collection*. Sand castles may be able to hold up frisbees in ways in which other collections of sand may not. Absent L2-*configuration*, L1-*collections* lack those powers. Bowling balls roll; their components in other configurations cannot. L2-properties do *not* obtain at L1; nothing about *resultant* L2-properties is mysterious. The mere location of each L1-constituent relative to the others lacks the causally-holistic-binding effect that results at the macro- or L2-level, as with perception. The holistic macro-element is *not* merely the collection of atomistic coordinates for L1-constituents.

Since causal/functional powers increase with configurational complexity, it is pseudo-emergentism. Just as the hard *determinist's actualism* is self-contradictory, science is up to its ears in pseudo-emergents and top-down causation (Kuppers 1992, Kauffman 1995, Gilbert and Sarkar 2000). Recall that CON seems to assume, but not to argue for, incompatibilism, as if non-autonomism is entailed by determinism, if not one of its axioms, as opposed to an interesting consequence of auxiliary arguments from determinism (Levin, 2003). Similarly, reductive physicalists often seem more to assume than argue for anti-resultantism, as if this were also entailed by physicalism, if not one of its axioms, rather than an interesting consequence of auxiliary arguments from physicalism. Just as determinism

does not *entail* hard determinism, physicalism doesn't *entail* hard physicalist non-resultantism. To assume otherwise is circular.

Just as CON wrongly interprets *Fixity* as axiomatically ruling out counterfactuals and *ability to do otherwise* as involving strong possibilities, hard physicalism wrongly interprets *the mental* as nonphysical and *the closure of the physical domain* as axiomatically ruling out the mental *qua nonphysical*, and so as ruling out resultantism. But if mind is thoroughly physical, then a robust downward mental causation and pseudo-emergentism are perfectly compatible – 'soft physicalism'. What constitutes most of L are special science laws that locate causation at the macro-level, either as lateral or downward causation, e.g., psychological affects hormonal (Sapolsky, 1997). With greater complexity comes *resultant* causal/functional powers: Life, homeostasis, and self-locomotion are such powers. The more complex, the less they are *directly* determined by their constituents and the more by their configurational properties.

With multiply-resultant holistic systems, the macro-properties may be metaphysically irreducible, e.g., the binding/unity of perception. But we may divide the problem of emergence into a hard and an easy one, ala Chalmers. The hard one serves as the repository for the metaphysical issues; we can solve the easy one by identifying the causal/functional properties that result at L2 from L1 objects, properties, etc. For easy autonomy purposes, the easy problem of emergence validates our notion of pseudo-emergence. The pseudo-emergentist, then, can take advantage of whatever the emergentist proposes, but in pseudo-emergent form.

The easy problem of emergence is to trace what novel higher-order properties are generated by what combinations and arrangements of older lower-order ones, while the hard problem is to state just what the metaphysical nature of the relationship is between the new

and old properties. The former amounts to offering a causal/functional analysis of the ways in which the old properties give rise to the new ones; the latter amounts to a metaphysical/ontological analysis of these cross-level relationships. We can ignore the latter here. Consider the Brussels-Austin nonequilibrium statistical mechanics model:

> If a system of particles is distributed uniformly in position and momentum in a region of space, the system is said to be in thermodynamic equilibrium (like cream uniformly distributed throughout a cup of coffee). In contrast, a system is far-from-equilibrium if the particles are arranged so that highly ordered structures appear (for example, a cube of ice floating in tea). (Bishop, 2002, p. 120)

Nonequilibrium statistical systems are identified by the presence of the following:

> large number of particles, high degree of structure and order, collective behavior, irreversibility, and emergent properties (*id.*).

The explanatory elements of conventional physics models are *particle trajectories* in physical systems, from which system behavior is derived and *reversible*, i.e., previous states are deducible from current ones. But the explanatory elements of Brussels-Austin models are *distributions*, i.e., *arrangements* of particles, and are *irreversible*. *Complexity matters!*

Bishop argues that features of this model support genuine macro- or system-level indeterminacy, without the difficulties found in conventional quantum mechanics.

> This implies that a system acting as a whole may produce collective effects that are not reducible to a summation of the trajectories and subelements composing the system (Petrosky and Prigogine 1997). The brain exhibits this type of collective behavior in many circumstances (Engel *et al.*, 1997), ... and this approach offers both an alternative for exploring the relationship between physics and free will and a possible new source for exploring indeterminism in free will theories. (Bishop, 2002, p. 121)

The claim that macro-indeterminacy is a feature of nonequilibrium systems like the brain would defeat CON and the 'cancels-out' bluff. Here is a respectable model of complex-systems-level *emergence* in contemporary physics. For easy-problem purposes, it is enough to appeal to a pseudo-emergent, pseudo-indeterminist version of the Brussels-Austin model.

Pseudo-emergentism is innocuous, so we may say that metacausal agents are complex nonequilibrium systems, whose components are nonequilibrium systems, e.g., consciousness, intentionality. The dialectical superiority of easy-problem/pseudo-version approaches may be applied to all the issues of freedom, e.g., reasons-explanations, self-formation; we may place all the hard ones aside, and retain their strongest features in *easy, pseudo*-forms. The easy-problem/pseudo-version approaches are more data-driven, evidentially basic, and less metaphysically-speculative compared to all the hard/non-pseudo-version approaches.

Thus, the causal/functional properties of autonomous systems are the complex evolutionary powers identified by the metacausal model, i.e., metacognition/metaconation. While Dennett (1984, 2003) has contributed to the idea that freedom is an evolutionary phenomenon identified with functionally complex meta-abilities, he does not say enough; neither does Frankfurt (1971), whose hierarchical account of the will and of freedom informs mine. But Frankfurt rejects the *causal* factors needed for the easy problem of *autonomy.*

Metamental abilities increase with primate development and growth from infancy to adulthood, peaking in counterfactual, prudential, and other imaginative reasoning. Off-line functioning involves the evolved ability to take cognitive/conative systems off-line. Such organisms are *freed from hydraulic, push-model forces that move organisms lacking these abilities.* Thus, they have *contra-hydraulic causal power,* a *natural* but *relative* form of *contracausal* power. For they can act contrary to hydraulic causal forces operating on them. As with the other pseudo/easy strategies, we need concern ourselves only with *easy* contracausality, *pseudo*-contracausality. Persons are pseudo-contracausal enough for easy-autonomy purposes.

The hydraulic momentum of on-line causation is not escaped by *merely* going off-line. The annals of meditation lineages are rich with phenomenological generalizations on

the discipline required to tame the *wild monkey* or *mad elephant*, metaphors for the ordinary mind. Western theology is replete with its version of the struggle, but characterized in polarized – if not Manichean – terms as the battle between angelic spirit and animal flesh.[90] The *Yoga Sutras* (Patanjali, 1953), *the* authority on yoga philosophy, revolve around the premise that *meditative union* results from the *control/cessation of mind-waves*.

Asian attention-training arts have levels of discipline we associate with athletes, and studies show their adepts score high on instruments that measure the effects of their efforts (EEG, EKG, etc.), revealing an extension of control into the autonomic nervous system inaccessible to most. In biofeedback, somatic information is looped to consciousness; by trial and error, agents learn to alter these stress indicators. The otherwise-uncontrollable adrenal-response, part of an on-line fight/flight S/R sequence, may be controlled by biofeedback. An agent takes that process off-line repeatedly increases conscious control over unconscious functioning. What brings the process under control is somatic information being fed back to consciousness, and practice.

Dreams, imagination and like states are intuitively off-line, as are language and simulata. A simulator who can entertain belief/desire inputs and choice/action outputs *without acting on them* instantiates a novel form of control over one's I/O. Biofeedback, meditation, language, simulation, and other evolutionary boons are disparate, but share in *off-line function*, *metacognition*, and *metacausality*, and so have common causal/functional features that reveal what autonomy is. I defined autonomy as strong/narrow metacausation. The weak sense involves any metacognitive causation, and the strong sense involves only

---

[90] Ainslie (2001) includes theological lore in "the data" for a theory of the will, for it is the richest source of folk psychological and phenomenological evidence on akratic features of will. The annals of Asian meditative traditions, likewise, count as a rich source of data for our theory.

top-down causal control of higher- over lower-order sequences. Since strong metacausation is *inherently causal*, there is every reason to think it is compatible with determinism.

Compare our model with Frankfurt's. For him, Agent's desires about other desires are 2nd order. The junkie has a 2nd-order desire not to act on her 1st-order desire to take drugs. When she succeeds, she acts freely; when she fails, she displays weakness of will. But unless she engages in downward causation, her success is not attributable to her 2nd-order desire. But someone indoctrinated with religious values against sex may have a 2nd-order desire not to act on 1st-order sexual desires, succeed at abstinence, and yet not make for an intuitive instance of freedom.

As Frankfurt pointed out, contrary to Hume, animals, young children, and other models of unfreedom typically have and are moved by primitive 1st-order desires; free agents typically have and are moved by meta-desires. Though *tacitly causal language drives many intuitions in Frankfurt's analysis* of the effectiveness of 2nd- over 1st-order desires, he *explicitly eschews* its relevance (1977). But not all metadesires are strongly metacausal.

Hierarchical *intentional* accounts are subject to regress problems, e.g., 2nd-order compulsion. The metacausal theory has no regress problem. As a *causal* hierarchy, it is different in kind from intentional hierarchies, so it is not subject to problems endemic to intentional hierarchies. Since we define autonomy as *downward* metacausation, and compulsion involves *bottom-up* causation, there can be no metacausal compulsion. The top-level in a metacausal hierarchy is defined as conscious, and there are brute cognitive parameters for an empirically possible state of consciousness.

> Finally, given that we are discussing a type of neural organization, there need be no undue concern regarding the classical problem of ... the infinite regress. Given that one part of the nervous system has the function of monitoring other processes, there is no difficulty in principle in having hierarchical levels of monitoring.... The limitation would not be one of logic but of diminishing gains in processing capacity,

which would rapidly lead to regression if infinite. (Weiskrantz, 1992, p. 197)

If consciousness is not narrowly meta*intentional* as opposed to broadly meta*mental*, then conscious states may be different in kind from lower-level states. The causal factor common to metamental states is that they are endogenous, mind-to-mind – not world-to-mind – engaged. Autonomous acts have endogenous metacognitive etiologies, authored or controlled top-down from the conscious level; *ergo*, there cannot be metacausal compulsion.

Autonomous action need not *be* off-line in the *occurrent* sense, so long as Agent *can* go off-line. The will is heterogenous, so each counterfactual ability adds to its hegemony. Mundane actions such as itch-scratching need not be consciously engaged, and need only involve counterfactual metamentality to count as autonomous; e.g., *had* Agent had meta-volition against it, he would have refrained.[91] Any cognitive/conative ability that can operate on itself involves feedback-looped off-line mechanisms functioning at a remove from the hydraulic causal nexus and makes *guided control* of response – a key feature of autonomous functioning – possible. Guidance is causal power to regulate the causal stream between I/Os to the sensory-motor system.

> Control is a topic of much discussion in the literature on autonomy. Sometimes it is claimed that agents have no control at all if determinism is true. That claim is false. When I drive my car (under normal conditions), I am in control of the turns it makes, even if our world happens to be deterministic. I certainly am in control of my car's movements in a way in which my passengers and others are not. A distinction can be drawn between compatibilist or "nonultimate" control and a species of control that might be available to agents in some indeterministic worlds – "ultimate" control. (Mele, 2003, p.1)

Mele cites Fischer's distinction between *guidance/*regulative-control (1994, pp. 132-35).

Fischer's guidance-control is Mele's *nonultimate*, compatibilist control illustrated in

---

[91] Many uses of counterfactuals have been entertained in the hierarchicalist literature; see Shatz (1986) for a critical review. But these affect only *metaintentional* accounts.

the driving example; Fischer's regulative-control is Mele's *ultimate* control. It is considered ultimate because if determinism is true, there are no strong alternatives or divergent cosmic reruns, so we cannot steer the course of our actions/events in a direction other than the one entailed/caused by ancient prenatal conditions, though it is we who do steer its *actual sequence*. Mele says we might possess ultimate control in an indeterministic world, for if in such a world there are strong, non-actual-sequence-type, alternatives lacking in deterministic ones, we could steer the stream of our actions/events in ways not constrained by Past-/Laws Fixity, which leave open only one path.

Fischer argues that while we possess weaker guidance-control, determinism rules out stronger, ultimate, regulative-type control. Fischer apparently accepts the CON equation that if determinism is true, there are no possible worlds other than the one determined by P_0&L, no alternate possibilities, and so agents cannot drive their cars in other directions – they *guide* the direction they go in, but cannot *regulate* it. This conflates ability to drive otherwise – given different motives – with Fixity-violating ability. Absent CON, the idea that agents are not what regulate/shape the stream of their actions is unfounded. The guidance/regulative-control distinction is unmotivated, for we don't need the kind of control CON says we lack – to produce divergent reruns – to have ordinary causal powers.

Agents possess *pseudo*-regulative-control when they satisfy PAPW2. CON does not undermine *this* control, sufficient for autonomy: An *easy-problem* solution need only specify *weak* control.[92] Thus, we have *pseudo*-regulative-control, absent CON, since our deliberative-mechanisms access *pseudo-randomness* (Dennett, 1981, pp. 294-99) and, since

---

[92] Fischer makes a similar case for what I call "pseudo-responsibility" where PAP fails, e.g., Frankfurt-cases, based on weaker control, but doesn't see the option for weaker autonomy. Fischer's arguments for guidance-based responsibility – the "semi-" in semi-compatibilism – can be modified in a weakened version of autonomy, thus *full* compatibilism, as in my account.

Fischer/Mele agree indeterminism permits regulative/ultimate control, pseudo-randomness permits pseudo-regulative-control. Causal/counterfactual control is pseudo-ultimate or pseudo-endogenous enough, yet determinism-friendly. If it turns out that we possess *non-pseudo*-abilities, we cannot be harmed thereby, but it is easy to prove – and enough for the easy problem to show – that we do possess all the *easy* or *pseudo*-abilities needed for a robust autonomy.

The key to the metacausal theory is its *easy/pseudo*-style analysis of configuration-level, emergent, causal/functional guidance/*regulative-control-generating* meta-powers. These are the self-system's abilities to cause/initiate/originate/guide/shape its own I/O mechanisms. At the inputs-stage are exteroceptions, proprioceptions, beliefs, desires, and apprehensions of RFAs.[93] At the operations-stage are deliberating, assigning weights, predicting, simulating, counterfactual reasoning, etc. At the outputs-stage are decisions, act-initiations, exertions of will, resistance against temptations, omissions, etc. These divisions are not exact. The operations stage may contain simulations of all three stages, for self/others in any combination. If Ainslie's model of will and self is correct (2001), agents engage in some form of future-self simulation in all deliberations, irrespective of whether the 'self' is integrated.

Perception of danger is part cognitive, part conative – survival-orientation. Neely (1974, p. 47) says *subjective* reasons matter, not reasons "out there" (*cf.* Bond, 1983; Mele, 1995). For, were Crito's reasons valid, Socrates' remaining in prison couldn't be free. But subjective reasons are also determined. Events, epistemic styles, preferences, etc. determine

---

[93] LTM shapes input by the unconscious selectivity of attention, what is stamped as salient before consciousness arrives, as the fight/flight response identifies what is survival-salient before cognitive processing feeds the information to consciousness (LeDoux, 1997).

the reasons I apprehend subjectively. These feed my deliberation, which feeds my decision/action, which effectuate a change in the state of the world (Levin, 1979). Any interconnection may succeed/fail. Successful control by Agent at *every stage* in the I/O process is coextensive with *total* or *ideal* autonomy (Mele, 1995); noncontrol at any stage involves *some* privation. *Maximal* autonomy is *not* necessary for responsibility or personhood, though metacausal control over *some* subsystems of voluntary behavior suffices for minimal personhood, responsibility and autonomy alike.

Metacausal powers are arguably conditions for the possibility of autonomy, but I do not adduce them by means of a transcendental deduction, but by means of data-level, easy-problem-oriented observation/analysis. This runs counter to a strong tradition in which the key adequacy condition for freedom is responsibility (e.g., Kant, Campbell, Frankfurt, Watson, Wolf, F&R). But conditions of responsibility (axiology) and autonomy (metaphysics) are not coextensive, much less synonymous, given the plausibility of normative pluralism, even if they are mutually informative or even conceptually interdependent. Thus, I may be deeply *causally* responsible for X-ing and hence *morally* responsible on our *desert*-theoretic view of responsibility,[94] but not on a Martian blend of utility and, say, some alien value, or a Venusian view that requires maximal metacausal control for full responsibility. Most think we are responsible only for things for which we *deserve* to be held responsible, as opposed to what will bring about *utility*.

The metacausal conditions for autonomy are pseudo-emergent abilities that solve the easy problem. Their description is sensitive to the literature on control, on external manipulation, and on Frankfurt-cases, and they handle the problem cases better than the

---

[94] Some adopt a utility view, believing CON undermines desert, but CON is implausible. Only desert-based views respect the supervenience of normative on metaphysical facts.

competitors. I replace PAP by PAPW, a weaker, pseudo-version that removes the need for strong alternatives and is justified by the dialectics of the easy problem and by the work it can do with Frankfurt-cases.

PAPW1: Agents are *self-guiding* with respect to their X-ing *iff* they *would have* X-ed had they been able to not-X.

PAPW2: Agents are *pseudo-self-regulating* with respect to their X-ing *iff* they satisfy PAPW1 and wouldn't have X-ed had they not wanted to X.

PAPW is supplemented by PAPM, the metacausal PAP-principle, roughly, as follows.

PAPM: Agents are autonomous (and, to the extent responsibility rests on autonomy, responsible) with respect to their X-ing *iff* they satisfy PAPW and metacausally control both this X-ing and their acts in general.

PAPM is left intentionally vague here. Its vagueness is removed later, but its informality is justified by the inherently vague and heterogenous features of the data.

Just as the Frankfurt literature unanimously accepts that Intervener need not *causally interfere* in Agent's decisions to qualify for *control* status, *metacausing one's x-ing* need not require Agent to *causally interfere* in his decisions to qualify for *control* status. Metacausing is pseudo-regulative-control. F&R's case of regulative-control is of the "Instructor" whose car enables Instructor to override "Driver's" turns, etc. (1998). Instructor possesses this control whether or not he exercises it. If Intervener and Instructor have unexercised counterfactual control, then so do Agents, for they can intervene if their 1st-order interactions with the environment steer them in directions not endorsed.

In Mele's driving example, I argue, one metacauses his driving even when he exerts no efforts on the control instruments, so long as their operation is under his conscious control, and he satisfies PAPW; he would continue to exert no muscular pressure on these control mechanisms were he able to do so, and were he to want to do so, he would do so. To think he cannot is to embrace actualism and conflate unexercised ability with inability.

There need be no alternate futures/worlds for it to be true that if he wanted otherwise, he could/would do otherwise. He is autonomous while driving, *whether determinism is true or not*, if he satisfies these conditions.

We stated the conditions that constitute metacausing one's X-ing; they involve natural abilities. These are causal powers of the will, so it will be useful to *have* a theory of *the will* itself, something overlooked by most writers on *free* will. Ainslie proposes a theory of *the will* (2001), more general than *free* will, so Ainslie's account provides the larger structure of issues against which background a theory of freedom may rest. Ainslie calls "picoeconomics" the study of micro-microeconomics, the same sort of relations as occur *interpersonally* between self-interested agents, but at the *intrapersonal*, micro-microeconomics level ("picoeconomics") level *within* Agent. Ainslie argues that the will is a functional process involving inter-temporal bargaining among a population of motives in a disintegrated self, and he uses this model to explain prudential, akratic, and related intentional behaviors.

Toward this end, Ainslie utilizes the following strategy: He uses 11 forms of evidence, identifies eight attributes of will, deploys three thought-experiments to test his theory, and compares four theories of will with his own, relative to these other considerations. His goal is to show that his picoeconomic theory both accounts for the identified attributes of will and handles the selected thought-experiments better than the alternative theories of will do on these fronts. It is a concise and intelligent argument form, with many merits and few defects. I adopt a similar strategy. My theory of freedom presupposes a theory of the will that falls roughly under the rubric of one of the alternatives with which Ainslie competes, *viz*, the 'organ' model, so I disagree that his theory better accounts for the attributes of will or the problems of will as represented in his thought-

experiments, though I otherwise adopt much of his analysis.

Ainslie considers 11 forms of data to show that his model is better than the alternatives: behaviorism, cognitive psychology, economics, philosophy of mind, psychoanalysis, bargaining research, chaos theory, sociobiology, neurophysiology, theology, and "folk psychology". I agree with Ainslie on their relevance to an account of the will, though I cannot repeat his reasons here. I add to his list. I draw evidence from attention training, biofeedback, evolution, computer-science flow-charts for on-/off-line functioning, developmental psychology, athletics, and related areas involving normative attributions regarding amoral notions of causal authorship.

Extending folk psychology, any human activity is relevant if it involves voluntary behavior. The ubiquity of the will in human life entails that there will also be found behavior-evaluating attributive practices, norms, principles, discourse, and a body of accumulated wisdom in every significant domain of human activity. This is the ubiquity of "the will", rather than "the voluntary", since in some akratic behavior it is unclear whether it is fully voluntary.

His eight attributes of will involve: its being a force distinct from the impulses with which it is engaged; its apparent application of its strength to the weaker side in a conflict; its action to unify actions under principles or ends; its strengthening by repetition; its asymmetrically greater sensitivity to weakening by nonrepetition; its nonrepression and nondiversion of attention; its resolve not depending on single choice except in special cases; and its variability with respect to failure in one sphere affecting failure in others. Other attributes need to be added to Ainslie's list.

One is a reflection of a key attribute of mental states identified by Searle (1992), namely, the fact that it is in principle accessible to consciousness. Two major, related

additional attributes of will are mirror reflections of key attributes of consciousness identified by Chalmers (1996, 1997), namely, the ability to verbalize any conscious state and to act on it. Thus, any act of will is one in principle capable of being accessible to consciousness, verbalized, and acted upon by will itself. Thus, any act of will is capable of becoming the lower-order content of a hierarchical will. Of course, it is a contingent, zoological fact that not all animals are metacognitively sophisticated enough to have the capacity for these higher functions or to develop them through fortuitous experience. Shatz (1986, pp. 465-66) says animals could have a higher-order preference, and that this undermines hierarchical accounts, but not our account.

The three 'thought-experiments' he identifies involve: Kavka's problem, free will, and Newcomb's problem. Each exposes assumptions about intentions. Kavka's and Newcomb's problems are logically equivalent.

Kavka's problem involves the issue of whether genuine intention is something we can move about at will, absent a forecast of whether the intention will be carried out. Consider an offer to receive a large sum for forming the intention, but only the intention (detectable by a sophisticated brain scan), to drink a noxious fluid. A similar case is Elster's (1983), where Ulysses knows the Sirens will overpower his intentions if his intentions are unaided by intention-overpowering prophylactics, such as his ear-plugged crew's binding him to the mast. If intention requires realistic forecast, then our toxin-consuming-intending test subjects cannot intend to drink the toxin, and Ulysses cannot intend to sail past the sirens unaided. (Ainslie, 2001, p. 126)

The free will problem as he sees it is the issue we grappled with in chapter 2. We will not address Ainslie's 'solution', for this *problem* was rejected with CON, and his solution is thin, i.e., picoeconomic determinism is consistent with determinism.

Newcomb's problem involves the thought-experiment in which a choice will be rewarded based on correct prediction in one of two ways, such that if Agent is correctly predicted to select two boxes, then box A will contain $1000 and box B will contain nothing, and if Agent is correctly predicted to select only box B, then box A will contain $1,000 and box B will contain $1,000,000. If Agent could get the $1,001,000 somehow, that would be optimal; the next best thing is the $1,000,000. If Agent fully intends to select only box B, then the prediction will function to produce $1,001,000 in the two boxes, but if Agent could, after the prediction, select both, she would get the optimal amount. Agent cannot do this unless Agent is capable of changing her mind at the last moment, but to the extent that this is her propensity, the perfect prediction should include it. But there is no reason why Agent should not try for both boxes.

These thought-experiments are designed to isolate features of will such as intention, prediction about intention, and the link between intention and action, and to help us sort out our understanding of the will and intention thereby. I appreciate the utility of requiring an account of will to be able to solve known problems of will, particularly the differences between free will, free action, compulsion, weakness of will and other privations of will, e.g., catatonia, and perhaps the normative differences that dovetail these, but I do not see the special relevance of the Kavka or Newcomb problems as tests of adequacy conditions on a theory of will in general. These two problems seem to be highlighted because Ainslie's theory proposes clever solutions to them, but that virtue doesn't translate into an adequacy condition on *any* theory of will.

Ainslie's solution is his picoeconomic inter-temporal, intrapersonal bargaining model on which a person is the intrapersonal equivalent of an interpersonal bargaining unit or marketplace of competing interests. Just as a stable market economy emerges by

sociobiological evolutionary processes, Ainslie proposes a stable internal economy is forged in the free market of a person's competing interests, which have in common only the fact that they share in the operations of the same organism or body. As such, Ulysses-like agreements among the 'crew' emerge naturally from their collectively rational self-interests the same way other facially altruistic arrangements do on micro-, economic, and macro-levels. That is, they develop what Dennett (2003), following John Maynard Smith, calls "evolutionary stable strategies" or "ESSs" (p. 149) from out of "benselfishness" (p. 194), the sort of enlightened self-interest Ben Franklin is said to have exhibited in his remark to John Hancock at the signing of the Declaration of Independence, "We must indeed all hang together, or, most assuredly, we shall all hang separately" (p. 193). I find Ainslie's proposals for these problems rich, but unconvincing in that they involve eliminativism about personhood, which he reduces to an internal marketplace of motives.

More pertinent are problems of will connected with its range, from full functionality and maximal autonomy to total privation: chronic motivational impotence, e.g., catatonia. Compulsion and catatonia mark the lower limit on the spectrum of will, and autonomy the upper limit; the Kavka/Newcomb problem is a special case situated in the mid-range of the spectrum. It would be a mistake to rest the plausibility of a theory of will, therefore, on how well it handles these problems. If there are logical impossibilities associated with certain self-indexing *predictive* paradoxes, then we just lack those abilities. Since most of what we do does not involve these putative abilities, nothing is lost. I have little motive to address the Kavka/Newcomb problem here, except to merely acknowledge them as special instances of the prediction paradox that affects soft determinism. Newcomb, at any rate, is often presented as a predictability puzzle, and this is not an intrinsically-deterministic issue. These puzzles are unnecessary *as tests of adequacy conditions* – which is the only reason I have

drawn in Ainslie's analysis to begin with.

The four alternative accounts of will Ainslie contrasts with his own are: the null model, the organ model, the resolute choice model, the pattern-seeking model. The *null model* sees the will as superfluous relative to the impulses which compose it: There is no will apart from the collection of conations. "The will" is analogous here with Hume's "self", a linguistic convenience – the "I" of speech. The *organ model* sees the will as a motor faculty analogous to an arm, something distinct from its components though made up out of them the way an arm or organ is made up out of its components. The *resolute choice model* says that the will acts by inhibiting reconsideration of plans; it is a selective function on elements which, once selected, are protected by the will against competing interests. The *pattern-seeking model* sees the will as an innate aversion against breaking up long-term motivational patterns for the sake of short-term ones, so it may be characterized by its habituated tendencies over time.

The metacausal theory resembles the organ model, but I resist the rubric as a set-up for a straw man. The metacausal theory subsumes the other models, for it views *all* features of the will under an all-inclusive spectrum of will; I call this the "subsumption thesis". What is true of behaviors exhibited clearly under one model of will may not be true of behaviors described well under another, but this is merely a function of these behaviors falling at different points along the spectrum of will. Models for behaviors in one range of the spectrum make sense for those behaviors, but not for others. The metacausal theory explains all of the behaviors that constitute the ubiquity of the will and its heterogeniety, i.e., its irreducibility to one motive-type, e.g., hedonism or libido, and why failing models are intuitive: They capture features of will in their ranges along the spectrum.

Hedonism explains most animal motivation, but in human psychology attempts to

reduce all motives to one type are not data-driven, and so are empirically implausible. Ainslie (2001) presents the utilitarian version of hedonism as an all-encompassing explanatory model, one that comes close to being sound for most behavior. But minimally data-driven attention to the data simply doesn't support any all-encompassing or monistic theoretical views of human motivation.

The major problem with Ainslie's model is its view of the person as a disunified collection of competing motives. A condition of adequacy on a theory of will is that it comport with basic intuitions about personhood. Ainslie's model doesn't, not that this is devastating, and not that disunified models of the self are contradictory or absurd; Buddha, Hume, and others have endorsed them with some good reasons, and they may be true, but they are too revisionary.

If freedom can only be saved by losing our selves, it is no easy triage. It is better to save self and autonomy; these are intimately linked (Frankfurt, 1971). The meta-will is a *sine qua non* of selfhood. If there were no agent authoring actions, but only an agreement among a temporally-extended population of competing interests within a mind, then whenever the population reached a different agreement – temporally extended or momentary – or experienced significant change in its membership of interests, the person would count as a different person. But if there is no entity that *endures* change in its motive-set and controls the behavior of the motive-population, no agent apart from the behaviors of the collectivity or motives "market", then who/what is responsible?

Were Ainslie's agent-eliminating model correct, it would be wrong for person P2 at T_2 to be held responsible for what P1 did at T_1, even if P2 occupies at T_2 the same body P1 occupied at T_1, so long as there is a different motive-set. So, P2 need not repay P1's student loans, since P2 no longer shares the interests that prompted P1 to undertake education. Here,

if anything is to even resemble an agent, were Ainslie to construct some functional equivalent, it would have to be the sort of entity constituted by the *long-term* interests of the motive-set. But if that were so, then the 'agent' could not be responsible for akratic behavior. Perhaps the interests that broke off from the agreement and gave into akratic forces would be specifically responsible, but the only way to hold specific motives responsible without punishing the entire person/population would be to engage some sort of selective behavior modification. And that seems highly implausible, not only pragmatically, but conceptually. That would be analogous to the not-long-forgotten practice of severing the hand of a thief, or removing some other offense-committing organ. The problem is that many non-offending members of the motive-population need to use these organs too.

This atomistic motivational model would also make it impossible to test whether behaviors are voluntary or causally controlled. For motivational agreements cannot be run coherently through the counterfactual principles – PAPW/PAPM – used to determine voluntariness/control. It would be impossible to run the tests in a way that captures intuitive judgments on bargaining agreements that vary over time and aren't responsible for short-term behavior. Thus, despite its rich ideas, Ainslie's model cannot be adopted in its current theoretical dress. Ainslie's work is best viewed as in progress; its goals/methods are inadequate for *our* purposes, but it is revealing in terms of depicting many features of motivation that need to be taken into account in any theory of will, thus informing our set of adequacy conditions.

Ainslie's model is motivational/conative; Fischer's is rational/cognitive. Each represents flaws in the other: Judgment alone, Hume insists, is impotent; motive alone, Kant insists, is blind. Clever concatentations overcome their deficiencies; e.g., Ainslie's model shows how unity of rational/utilitarian purposes emerges from a population of self-interested

motives. Fischer's model absorbs wants as RFAs, and accounts for the link between reasons/actions by treating actions as reactions to reasons, and thus undermines the will/judgment dichotomy. But *neither wants nor judgments* are actions, though *choices are.* The *will* is distinct from both, but is more wants- than judgments-theoretic, since wants move us but reasons alone cannot, and the will moves us. But it is distinct from wants as *an executive power with control functions on and over wants.* It also plays a role in propositional-attitude formation – there is *some* doxastic autonomy, limitations noted, contrary to Descartes. The will has greater *immediate* power over wants than judgments, but since it has executive power over both, it is distinct from, though shaped by, both.

Selection of the 'volition' from among its input is an *action* the will *performs*, and so falls on the *executive, motor-theoretic, reactive side* of the equation. Choice – and the willful initiation of teleological action that it originates – is distinct from motivation/cognition, though informed by both, as even libertarians hold. Pseudo-regulative-control enables Agent to alter his behavior regardless of how conscious or deliberate its *origins*. (The President, in *this* version of the popular analogy, need not author the legislation he endorses or vetoes.) The *character* of *agency* seen in this control is primarily *executive*, causal-originative power.

Some self-regulate; others don't. One reason reason-responsiveness is insufficient is that not all who control their wills are reason-responsive. But note that wants-responsiveness is not equivalent to reason-responsiveness, for a want is not necessarily a reason: A push-theoretic *force*, e.g., proprioception of hunger, is not a reason by itself, though it propels (conation) action. To convert a want to a reason, what is necessary is the formation of a pro-

attitude or judgment of the form that the want is worthy of pursuit.[95] Thus, wants need judgments to become or count as reasons. But propositional *contents* are transformed by the formation of pro-/con-*attitudes*. Thus, even pro-attitudes or judgments that convert wants into reasons involve the will: Forming a pro-attitude toward a want is a meta-will activity.

Reason-responsiveness is a condition of moral – not causal – responsibility. Fischer developed this model for purposes of semi-compatibilism: to account for the compatibility between determinism and moral responsibility while rejecting autonomy as incompatible with determinism. One who is aware of his actions and can interfere in them should his wants change has pseudo-regulative-control. A machine could have this power, but be incapable of recognizing *moral* reasons, a further condition F&R place on morally-responsible reason-responsive-mechanisms (1998), yet still be causally responsible for its behavior, so long as it is capable of sensing and altering its behavior. The same holds for a sociopath. *Free will is thus a motor-nerve-involving, self-regulating, executive causal power over one's actions.* The more sophisticated one's motivational systems and the more intelligent, the more options to select from, and the more extensive the range of one's will becomes. Beings with lesser motivational sets and minds have less sophisticated/intelligent wills, but these are extrinsic features of will.

A nickel is non-zero currency, as is a fortune. They differ, but both are money. Likewise, a being with few wants and sensory-motor skills, say, a rat, has a small will, and a being with many wants and judgments and refined sensory-motor skills has a greater amount/range of will. Free will, then, is more than just will, but the ability to control one's will, as Frankfurt almost said. It is an executive power one can exercise over one's wants,

---

[95] On judgments that a want involves a good to be pursued, see Bond (1983), Watson's Platonic version (1975), and Stump's reasoning (1993), citing Aquinas; *cf.* Nagel (1970).

judgments, and even itself, as will inclines in a direction one wishes to alter. That executive meta-will function is self-regulative causal control, autonomy.

Ainslie's model is interesting because it instantiates an empirically possible soft determinism, as his brief discussion reveals, and it suggests what an adequate theory of will should do. Extracting from his strategy, an adequate theory of will should:

(AC1) account for the known attributes of will, and
(AC2) deal with known problems of will better than the alternatives do.

In my view, an adequate theory of the will must do at least two other major things. It must:

(AC3) comport with a theory of the self or personhood, and
(AC4) comport with a theory of responsibility.

Levin's conditions of adequacy are for a theory of *free* will (1979); they are adopted in our theory of *free* will as well, together with conditions AC1-AC4. Ainslie mentions eight attributes of will, but there are others, e.g., its ubiquity, heterogeneity, accessibility to consciousness, reflexive accessibility to itself, susceptibility to reason, privation relative to diminished intellect, pivotal role in artistic or athletic accomplishments. The Kavka and Newcomb problems are inessential to an account of the will, but there are problems he doesn't mention, e.g., privation cases, those in Frankfurt-style cases, and machines capable of Ainslie's intertemporal motivational bargaining.

A theory of will also ought to account for: phenomenological intuitions, such as its being up to Agent to manipulate the will one way rather than the other; the grounding of normative attributions in these powers of the will; the grounding of exculpation in the absence of these powers; the distinctiveness of these powers to persons; and those intuitions connected with Levin's (1979) adequacy conditions.

The metacausal theory rejects the null model on zoological grounds. Sentient beings comprise the class of beings with will. This class is problematic only at its penumbral edges.

If biologists held that cellular photosensitivity is a primitive form of stimulus-*cognition* and thus of *sentience*, but tropistic response doesn't count as sufficiently *motor-theoretic* to house auto-motive response-*conation*, then only *most* sentient beings comprise the class of beings with any form of will; the converse, that all beings with will are sentient, is less risky. While paraplegics have sentience without motor-control, it is doubtful any will-possessing being can be fully insentient. Brain/behavioral research supports the claim that willful, motor-theoretic behavior is funneled through the consciousness centers of the brain for orientation purposes. Barring penumbral cases, the sentience-will equivalence holds.[96]

Behavioral evidence supports the view that the motives in primitive sentient beings involve attraction to life- and/or health-promoting pleasant-sensation-producing stimuli (positive reinforcement) and aversion to death- and/or illness-promoting pain-producing stimuli (negative reinforcement), and though there are penumbral cases between these and tropisms that only *appear* to involve hedonistic sentience, once we are clearly dealing with hedonistic sentience, we are dealing with the lowest form on the spectrum of will, and that is cognitive. The line may be less than straight, but we can draw it clearly enough: The will is thus *hybrid*, cognitive-feedback-looped-conation, inherently sensorimotor-theoretic, and thus inherently metamental, involving, as it does, two mental state – cognitive/conative – elements about the same object (telos).

There may be non-penumbral exceptions to the claim that primitive motivation is hedonistic, e.g., bees whose instincts propel them to death for the queen; the empirical question is whether these behaviors are a function of expected reward. Lab chimps self-

---

[96] Somnambulism and other exceptions are explainable on the basis of an error theory about blown neural fuses/switches for nocturnal akinesia. In sleep, motor-responses are taken off-line, or we would really run when chased in dreams. Sleeping animals often exhibit damped down motor activity indicative of fearful dreams. On studies favoring this view, see Winson (1997).

medicate on cocaine until they die. It is more reasonable to the pleasure- over the death-principle. More complex cases involve smoking, obesity, and other self-defeating behaviors tied to the pleasure motive. But all animal behavior seems hedonistically-oriented, reward-expectation-driven, and most of it is well-correlated with survival. To the extent that mankind is not equally constrained by the ordinary operation of natural selection, our behavior is exceptional regarding failure to exhibit survival-orientation, apart from its exceptional character on other grounds.

To the extent that a motive is a kind of "willing", a positive response or movement toward an expected reward as its goal, then, even the most elemental motive expresses *some* form of will; this is so also for any attraction or aversion, e.g., infant crying. On this analysis, one might hold that complex will reduces to details about complexity, so there is no need to treat the will as if it is a whole that is somehow separate from the sum of its parts, as if motives are the atoms and the will is just the collection of those atoms, or as if some of those atoms of will are less relevant or even 'less will' than others of them. But this misses a distinction. It threatens to collapse at the border between positive responses toward stimuli, which tropistic behavior involves, and attempts to repeat the stimuli, which minimally involve associative learning and characterize teleological motion, absent in mere tropisms. But assuming a clear line may be drawn to distinguish teleological responses from nonteleological tropisms, at the base-level of teleological motions are such beings as slugs, who apparently act on elemental motives, but lack any structure to their motives. The more complex the species, the greater not only the learning mechanism, e.g., more complex orientation-feedback mechanisms, but also the structure of the will. It is only as we move up the evolutionary scale of complexity to cases where there can be a conflict in hierarchically-configured will enabled by hierarchical *mind* that we are willing to make unhedged

ascriptions of will.

As Frankfurt says (1971), the child and animal act on the stronger of two or more motives, but only persons exhibit deep conflicts of will. Thus, though "will" *may* be applied inclusively to all forms of motivation, it is *more intuitively associated* with its *less inclusive* usage evidenced by most of Ainslie's attributes of will, which refers to a *relation* within a motivational set, and features of the relation are what account for the possibility of differentiating between intended/unintended actions, or behaviors that express true preferences and those that don't, even though in all four cases the behaviors involve motives. Thus, the *exclusive* will involves relations between sets and subsets of motives; thus, the null model is false.

The alternative requires that the will involves relations between motive sets/subsets. The will is structured and divided, whether it is an organ that selects or maintains the subset of motives, a function that resolves to maintain the subset in the face of reconsideration of members in the larger set, a tendency to maintain whatever pattern long-range reward-expectation has designated as the proper subset, or Ainslie's intertemporal relation among bargaining members of the larger set. On the metacausal theory the will is comprised of causal/functional relations between motive set/subsets, and it is consistent with the view of motivational atoms discussed under the null model.

The resolution to ignore reconsiderations of non-subset motives is like the tendency to maintain long-term subset motives, so they may be grouped together as instances of "subset individuation and/or maintenance criteria". Ainslie's model may be subsumed under this rubric. The individuation of a subset of motives provides the *content* of the will, but at any given moment Agent can endorse a single motive and make it his will, even if it runs counter to any and all indications that favor long-term subset maintenance (willed ignorance

of reasons for reconsideration, preference for long-term-pattern-approved subset members, subsets identified by intertemporal bargaining). Thus, each model identifies a unique ability of the will not only to individuate its contents by that model's criterial-selection process, but as a method of will-content-formation to override the other models' types of process for the formation of the will. Thus, willed-ignorance toward reasons-for-reconsideration is a process the will can deploy to subvert the tendency toward intertemporal bargaining or its tendency to maintain a long-term pattern. Thus, all non-organ models collapse into fingers on the more inclusive hand of metawill. The metacausal model absorbs every attribute of will and accounts for all intuitions associated with intentional action.

The understanding is a high-level faculty with no organ; though the parts of the brain that subserve it may be an organ-like scattered individual linked by neural networks, understanding is more a higher-order functional state of the brain than a mere constellation of neural parts that might be required for an 'organ' of ability, e.g., visual memory. By analogy, the will is subserved by brain parts, but it is not the parts of a neural network 'organ' that form the will, but the functional relations they subserve. Given brain plasticity, multiple realization, the variable nature of higher-order faculties such as understanding/will, and the heterogeneity of the will, 'organ' is misdescriptive. Though the supplementary motor area of the cortex has been identified as a candidate *seat of* the will (Northoff, 1999), it cannot *be* 'the will', given plasticity and neural looping through other brain centers. If *a mental state must be subserved either by an organ or by metamentation*, then the will must involve metaconations, since there is no organ of will.

If the subsumption thesis is correct, then the metafunctional character of the will (data) makes the metacausal model (theory) most appropriate. The meta-ability to activate a motive in one's overall motive set, for any or no reason, is what constitutes the will as a

functional power over and above, and capable of operating upon, all its components and their subsets. This meta-ability comes in degrees. The more *inclusive* "will" that applies to *any* motivation does not constitute "*the* will", the more *exclusive* meta-ability to act on or activate any motive. So, to reject exclusive will by reference to inclusive will is to equivocate.

The metafunctional model captures the experiential character of the will as an ability to do or not do what a motive impels one toward (Locke), what one pleases (Hume), what is rational (Kant), and to not do what one wants now because one wants something greater later (Ainslie). All these cases involve exclusive will acting on inclusive will, so *exclusive* will – *the will* – is metaconative; inclusive will is not. The heterogeneity/ubiquity of the will supports the subsumption thesis and the metafunctional character of *the will*, and thereby the metaconative – hence metacausal – model. Freedom comes in degrees, and its upper boundary is metacausal autonomy; weakness of will, conversely, is also graded and bounded.

As Frankfurt argues (1971), only beings capable of *conflicts* of will are candidates for weakness of will, and slugs do not seem so capable. If so, then we ought to say slugs have no exclusive wills, though they have inclusive ones. Non-structured wants are better called "wants", and "will" is better restricted to the narrow sense. Most animals possess wants, and thus some freedom of inclusive will, which is what Frankfurt has in mind about *freedom of action* (the horse wants to run to the left, and does); most of us possess some freedom of exclusive will (Frankfurt's free will) and plenty of inclusive will (Frankfurt's free action).

While there is a link between freedom of inclusive will and freedom of action, these may come apart. Fidelity to the data suggests inclusive/exclusive will both be treated as forms of will, for an animal may have inclusive wants and be free to act on them without doing so; its inclusive will is free counterfactually, for it can act on its wants, should it want to. Circumstances may be manipulated so that though the animal enjoys freedom of

inclusive will, its actions are all foiled. Frankfurt's taxonomy needs revision. Conceivably, a smart manipulated beast may develop a structured will by its learning associations from this conditioning process.

Animals have freedom of action, since they act on wants, but not of will, since they cannot alter their will (Frankfurt, 1971). Frankfurt should say they *typically* enjoy freedom of inclusive will, whether they also enjoy freedom of action. To the extent that maximal metaconative ability is coextensive with freedom, and restraint is metaconative ability, rats have some freedom of exclusive will, for they can avoid cats though there is rat food near the cats which attracts them.

Consider Ainslie's contrast between two models for decision making.

> Models based on wanting say that people weigh the feeling of satisfaction that follows different alternatives and selectively repeat those behaviors that lead to the most satisfaction. Models based on judging take a hierarchy of wants as given and focus on how a person uses logic – or some other cognitive faculty – to relate options to this hierarchy. (2001, p. 13)

Want-based models are "hedonistic", "economic", or "utilitarian"; judgment-based models are "cognitive" or "rationalistic" (*id.*, p. 14). The Socratic question about akratic behavior shows the dichotomous character of these models: Either judgment succeeds but will is weak, or, with Socrates, judgment fails, but perfect will is misguided.

I am concerned not with the Socratic question, but with want/judgment models. Wants may be viewed hydraulically, as "motions" or pressures built up in the organism and seeking release, as Hobbes, Freud, and others (e.g., McGinn, 2003) seem to view them. Most animals, many young children, and some adults seem driven by wants, but many adults seem guided by judgments. I place hydraulic wants on the *animal* side of the spectrum and nonhydraulic judgments on the *person* side, with overlapping in the middle range. Call this the "want/judgment animal/person spectrum". Restraint may involve some wants overriding

others, e.g., stronger occurrent or prudentially-ranked wants overriding immediate wants triggered by environmental stimuli. Restraint may also involve nonprudential or immediate judgments overriding immediate wants, e.g., judgments about the immediate consequential/moral implications of satisfying the wants. Competing immediate costs have competing hydraulic quantities, e.g., the rat's food/cat equation, but moral implications may also have greater hydraulic qualities.

Thus, one may treat want/judgment in terms of two-tiered hydraulics, qualities/quantities, akin to Mill's treatment of pleasure. Hard determinists would be sympathetic to this Newtonian mechanical view, but the plausibility of Ainslie's distinction between *push-pull-models* of motive-types undermines attempts to reduce motives to hydraulics. Some motives propel as a force from behind, a pressure that pushes us forward from the rear; others pull us forward with attraction toward desired goals. Hydraulic motives are push model, but most of what motivates people are pull-model motives, and these are not hydraulic. These are blended, and opaquely so.

Even on the hydraulic model, restraint (exclusive will) involves the novel causal power to disengage the evolved organism from its otherwise total (hydraulic) functional/operational connection with the "organism/environment field" (Perls, 1947). Organisms unable to disengage are "on-line" with the S/R stream that deterministically feeds hydraulic force (inclusive will) through the I/O mechanism; organisms that can disengage from the S/R stream of inclusive will put a causal/functional wedge between the ordinary S/R causal pairs, and are "off-line" from the hydraulic deterministic stream (inclusive will) that necessitates primitive behavior. This counts as a novel break in the chain of hydraulic determinism – pseudo-emergentism.

Even if restraint (exclusive will) can be reduced to purely quantitative hydraulic

want- or push-model terms, as opposed to qualitative, judgment- or pull-model terms about the good to be pursued, the off-line organism is at an advantage over the on-line organism, for the former has options the latter lacks. But if judgment-model restraint is not reducible to the quantitative hydraulic model, then it constitutes a greater off-line power than the more base, want-theoretic form of restraint, still off-line, exclusive will.

Absent reasons favoring one model, wants- and judgments-based decisions are distinct enough to place animality and wants on the low end of the spectrum, and personhood and judgments on the high end. Both forms of restraint empower their possessors with a degree of autonomous off-line functionality over on-line organisms, though judgment-theoretic restraint is a greater boon on the want/judgment animal/person spectrum.

A key reason that judgment-model restraint is superior to want-model restraint is that the understanding can reassign values to wants (inclusive will), rearrange the hierarchy (exclusive will). Want-model restraint is just a function of hydraulic (inclusive will) forces, characterized as *equal* in the Buridan case. And judgment-model restraint can bring considerations to bear on a decision that it expects to incline the will one way rather than the other, as in the pragmatic self-orientation of attention in attempts to form health-promoting beliefs; it can ignore considerations having an adverse expectation, as in the tenacious maintenance of a dogma or an unforgiving attitude, as with Goebbels (Stump, 1993); and it can call off or continue reconsideration of the factors as it sees fit, as Ainslie's *resolute choice* model suggests. Sophists' and Pyrrhoneans' abilities to juggle reasons come to mind. To the extent that it can juggle competing wants, reasons, values, and judgments, it has plenty of elbow room and off-line power – the resultant power to transmute hydraulic forces

striking the border in the organism/environment field.[97]

Autonomous agents can intentionally equalize the determinants of their behavior; they can reduce addictive responses by avoiding addictive stimuli, as Ulysses did with the Sirens (Elster, 1979). Analogously, Rawls' "veil of ignorance" strategy schematizes a technique agents can use to screen out undesirable influences (1971). Agents who wish to select principles of action can follow any number of analogues of this decision procedure to escape from what Kant calls the heteronomous causation of the will. Science may enable us to equalize what evolution has given us. Cyborg persons of the future can autotelically rewrite their genetic sequences and their neurosignature sequences at will from among a vast array of internally apparent programs, some involving random number generators, both pseudo-random and fully random. These beings have greater autonomy than we do, but we need not make it to the fully autotelic, autogenous cyborg level to see that we are much closer to that than we are to the animals, even now.[98] The more autonomous among us are virtually there already. Freedom admits of, and comes in, degrees.

This off-line power enhances the intensity and ambit of the will, which are in turn enlarged by repeated use, as a muscle is so enlarged (one of Ainslie's attributes of will, *supra*), and by any and all functions that engender restraint. Reflexive operations typically involving feedback-oriented cognition are restraint-empowering factors, as evidenced in biofeedback (Olton and Noonberg, 1980) and meditation (Csikszentmihalyi, 1991).

---

[97] See Perls (1947) for a rich analysis of the causal/functional relations at the "contact boundary" in the "organism/environment field". Based on earlier theoretical work in the *gestalt* theory of perception, Perls *et al.* (1951) identify attention and its ability to determine the figure-ground relationship or *gestalt* as key determinants of healthy organismic functioning.

[98] Ironically, these considerations suggest that the first *fully* autonomous beings will likely be futuristic hybrid (part-biological, part-synthetic) persons (cyborgs) or "wet" AI machines. Here is where one may appropriately embrace a *reductio ad absurdem* to the effect that one's model of freedom entails that machines can be free, and say "So what?" in response.

We possess novel judgment-theoretic abilities. We can engage deeply reflexive cogitations and re-valuational operations on motivational atoms and the hierarchical structures that distribute them within the valuational array of our awareness, deploy Buridan-case tie-breaking pseudo-random and reason-alternating operations to incline/disincline the will, and shift figure-ground configurations in our awareness. We can do all this while functionally disengaged or off-line from interactive involvement at the contact boundary in the organism/ environment field. Surely these are emergent causal autonomy powers, powers over the hydraulic determinism of mere animal wants and their on-line, immediate S/R functioning.

We can engage the hydraulic force of animal wants to overthrow reason, by guzzling down alcohol to fuel rage, so the will seems to have access to – and thus to be distinct from, to transcend – both wants and judgments. So, in its transcendental function the will involves a hybrid of reason and desire in that it is the *wants-informed (inclusive will) intelligent ability (exclusive will) to direct the attention*,[99] and thus to direct the figure-ground relationship, and so to move the organism in its total functioning, on-/off-line, *at will*.[100] Driving this whole process is the attention, consciousness, as depicted metaphorically in the role of the soul as charioteer over the horses of passion, both in Plato's *Phaedrus* and in the *Bhagavad Gita*.[101]

---

[99] On the control powers of attention, see Austin (1998), Benson *et al.* (1990), Chalmers (1996, 1997), Csikszentmihalyi (1991), Forrest-Presley *et al.* (1985), Goleman and Thurman (1991), Hamilton (1976, 1981), Hamilton and De la Pena (1977), Hamilton *et al.* (1984), Marcel and Bisiach (1992), Metcalf and Schimamura (1994), Nelson (1992), Olton and Noonberg (1980), Perls *et al.* (1951), Robbins (1997), Umlità (1992), Weinert and Kluwe (1987).

[100] "At will" refers to inclusive will in beings who possess exclusive will. Thus, it sounds odd to say a bird flies 'at will' (or 'freely'), since there is no contrast against the background of which its flying is distinct (compared to everything else it does). "At will", then, is a hybrid of exclusive and inclusive will that doesn't apply to beings with nonhierarchical wills.

[101] *Phaedrus*, 246a-b, 253c ff; see Zaehner (1966) for a good translation of the *Bhagavad Gita*. A similar model is in the state-soul analogy in Plato's *Republic* (1941): Reason informs will, which controls passion. This model is critically reviewed in Blackburn (1998).

Attentive consciousness – meta-awareness – holds the reins of exclusive will more than any part of Agent. LTM determines what elements of the perceptual field become figural in the figure-ground *gestalten* that emerge in consciousness. Exclusive will rankings cultivated over time into habits largely determine what LTM selects. LTM thus embodies the residual hierarchical structure of what Kane (2002b) and others call the self-system; Frankfurt (1971) and Perls (1947) might agree it is a system of identifications/alienations. Our voluntary behavior of yesterday forms our character of today and shapes what seems to arise in awareness spontaneously, not only in the selectivity of perceptual attention from the influences of LTM, but in the generation of the sorts of reasons that will count as "live options" (James) that we will consider when deliberating, etc.[102] Since each time we repeat an action we reinforce it (Ainslie), we are responsible for our self-formation, for what sorts of urges, ideas, etc. occur to us 'spontaneously', and partly for our brain structures.[103]

These are more a function of our *idiosyncratic* exclusive will patterns than they are of ancient causal streams that may have culminated in the raw materials that constitute the inclusive-will inputs that fed into our early exclusive-will-engaging behaviors. Since exclusive will brings so much more self-sculpting and inclusive-will-transcending power and largely determines what inclusive-will-type RFAs will even appear as triggered spontaneously by conscious experience, CON-ultimacy arguments fail on any close examination of the relevant details.

These intuitions about self-formation link with libertarian views of Agent as acausal

---

[102] Mele (2002, 2003) places indeterminism in the generation of deliberative reasons, but that would have little effect if what is identified as salient is a function of past performances; but *if it were effective*, it might repeatedly *threaten* to change Agent's character.

[103] Since there are neurosignatures created in the elastic brain from each repeated experience (Benson, 1996), we significantly create our own brain structures by our choices/habits.

or contracausal, but without any appeals to indeterminacy or any other unnatural/hard doctrine. The ability of the will to transcend both wants (inclusive) and judgments (exclusive) accounts for the phenomenology, as if it is beyond the causal stream altogether. Just as pseudo-random processes are virtually random enough, so too the transcendental nature of exclusive will is pseudo-acausal, and thus virtually contracausal enough. This power, though consistent with deterministic causation at more global levels, *seems* contracausal because it *can literally override the base hydraulic determinism* (inclusive will) that impels animals. Thus, *it is* contracausal *relative to hydraulic causation*, though only pseudo-contracausal relative to determinism *simpliciter*.

What is the *character* of this higher will, is it more want-theoretic (inclusive will) or judgment-theoretic (exclusive will), or something else? What distinguishes us from other animals in terms of will is our exclusive (hierarchical) will and their inclusive (nonhierarchical) will. Action involves teleological (wants-involving) movement (or restraint, however slight, in *intentional* omissions), distinct from the deliberative process, though guided by it. Aristotle's model of the practical syllogism, whereby the conclusion of deliberation is a judgment about the good to be pursued, which issues in the related action, is phenomenologically false. The will is capable of halting the action or executing the judgment *after* the decision is made. While reflection on the Kavka/Newcomb problem suggests intentions cannot be properly individuated independently of forecasts about the likelihood of their being carried out, in representative cases, decisions and choices can be defeated, though in some cases we view them as success terms.

Are decisions distinct from choices? Intuition differentiates them, though usage often joins them. I can study the menu and decide to order the filet of sole, but choose to change my mind in the presence of the waiter, having just seen something more appealing served at

the next table. Or I may be unable to choose it if the waiter says they have no more filet. These cases suggest decision is more tied to bringing deliberation to an end, but choice need not involve deliberation. "Choose a card from the deck" is a phrase that supports this, for there is no deliberation in this case; "decide on a card, any card from the deck" is grammatically awkward.

The difference between exclusive willings supports cases like deciding to vote Republican but choosing in the booth at the last moment to change one's mind and vote Democratic. Some cases turn on the point where intervention makes sense or not; e.g., "flicker-of-freedom" theorists argue that the slightest effort in the direction of a choice is a different choice than the one Intervener would bring about, were that flicker evident (Fischer, 2002). Choosing is about actual-sequence selection from alternatives and decision is more about the final ranking of deliberated options, which need not be instantiated to count as a decision. F&R's distinction (1998) between reason-receptive, -reactive, and -responsive behavior supports this. Deliberation is reason-receptive, decision is reason-reactive, and choice is reason-responsive.

In its highest form, exclusive will is the ability to transcend and equalize hydraulic and teleological determination, or both, and these are often opaquely-intertwined. This novel dual ability puts agents *causally/functionally* above both the hydraulic stream and the stream of rational compulsion, e.g., by truth. Compared to the S/R causation of on-line behavior, off-line powers are virtually *a*causal, and all such powers strengthen/extend the range of metacausal power. So, the power to use reason over want or vice versa, or any other form of restraint, adds to the virtual supracausal power of the meta-will, compared to the will of animals, just as the hydraulics of animal wants are a leap over the tropisms of plants, and the latter an improvement over the even more purely hydraulic causation of Newtonian

mechanics seen in the collision of two bodies. The metacausal will is, virtually speaking, uncaused, though only pseudo-uncaused. This view naturalizes the teleological rejection of reasons as causes put forward by libertarians.[104]

Decision may *sometimes* be a success term. I can choose the fish on whim, exhibited by the convention that I am willing to pay for it if it is delivered, only to be disappointed when the waiter returns with the bad news. *Here*, choice seems to involve a commitment-component related to a success-criterion, but both fail as success terms when I decide or choose to win the election, since it is not in my power to bring that about. I can decide or choose to run for office, even if it turns out that I fail to get the needed signatures. One can decide to spend the afternoon playing handball with a friend, but sprain an ankle on the way to meet the friend. Kant's reasoning about motives being crucial for moral evaluation is insightful to the extent that actions and consequences are partly outside the range of our full control, but erroneous to the extent that it ignores the fact that reasonable expectations or forecasts about success (Kafka/Newcomb, Ulysses) also play a key role in our evaluations of decisions, in addition to our evaluation of RFAs. For it is appropriate to hold someone blameworthy if we reasonably expect he should not have attempted an action with little likelihood of success, which entails that we judge him able to have perceived this and to have succeeded in doing otherwise.[105]

Wants and decisions differ, whether the latter involve success or not. Wants involve expectations of reward, often hedonic, and we associate hedonic rewards with the stimuli that have generated them, but there is a difference between *anticipating* that a pleasant

---

[104] See Collins (1984), Davidson (1968), Ginet (2002), Grice (1980), and Parfit (1986) on reasons as causes.

[105] Widerker's "PAE" principle (2002) formalizes a general version of this intuition.

experience will arise if one X's, the *yearning for* that sensation which *inclines* one to X, and being *actually driven* by that hoped-for sensation to do X. The latter may or may not involve bodily (proprioceptive) sensations of its own, as with hunger, lust, and other powerful emotions; these often have a phenomenologically-transparent hydraulic character, but not always a first-personally transparent one, as per Freud. Part of this is a hedonistic craving for a certain pleasant sensation, part of it is a judgment that X generates that sensation (and part a judgment that this sensation is a worthwhile pursuit), and part of it is the moving toward the good to be achieved by X-ing; again, part of it may involve proprioception of hydraulic force. It is complex.

Metacausal power enables Agent to rise above hydraulic/non-hydraulic will in an off-line meditation about RFAs and their goals, and it can value/devalue, activate/deactivate any motivational atoms/wants or RFAs/judgments, in transcendent *pseudo*-homunculus form, with all want/reason causal feeds so stripped of hydraulic force as if to be uncaused, though still causal. This accounts for the phenomenology involving S/R-free, hydraulics-free will, similar to God's acausal, transcendent freedom to choose from among possible worlds, as a prime mover unmoved. Contrary intuitions are explained by locating them on the lower end of the spectrum.

The metacausal model handles the want/judgment, reason/passion dichotomies and subsumes their features as evidence for its comprehensive explanatory power. It provides the causal/functional metamental architecture on which the principles of voluntary/guidance and determinism-friendly counterfactual/regulative-control, PAPW1 and PAPW2, and PAPM-enriched versions of the same, may be grounded. Let us turn to these.

Chapter Five: The Metacausal Principles, PAPW and PAPM

Let us begin with a review of PAPW1, the principle of voluntariness, and PAPW2, the principle of weak conditional ability.

PAPW1: Agent performs an action voluntarily *iff* in acting he *would* have so acted *had* he been able to do otherwise (had he had access to alternatives);

PAPW2: Agent has weak conditional ability in performing an action *iff* in acting he satisfies PAPW1 and *would* have done otherwise *had* he wanted to.

In Locke's secretly-locked room, Agent stays *voluntarily* though he cannot leave, under PAPW1, *iff* he would have stayed *had* he been able to have gotten out. But he lacks weak conditional ability, since PAPW2 is not satisfied: He wouldn't have left had he wanted to.

In a Frankfurt-case, Smith cannot vote Republican because Intervener is ready to block him from wanting to, though he wants to vote Democratic on his own. I used to think he fails PAPW2, and agreed with the majority that Smith *cannot* vote Republican, thanks to Intervener, but votes Democratic voluntarily, since he would have voted Democratic anyway, had there been no Intervener or implant. I was content to be able to differentiate PAPW1 and PAPW2 to account for the fact that he acted voluntarily while being unable to do otherwise. But a valid question is: *Would he have voted Republican if he had wanted to?* Like many, I thought "no", *he cannot want to*, for Intervener will not allow him to want otherwise, which is fine, for in otherwise analogous cases lacking Interveners, Agents satisfy PAPW2, and that suffices for non-phi-fi autonomy.

I have come to think the answer is only "no" *in the sense that he cannot want to*; whereas, if there is some sense, or some world, in which *he could want otherwise*, it is plausible that he would vote otherwise *if* in the actual world his wanting otherwise in the range of relevantly similar cases is *usually* effective in his doing otherwise. I have been influenced in this line of thought by Levin (2003), who argues that what the Intervener or

implant prevents is his *wanting* to vote Republican, not his voting Republican, in which case he might still have some weak conditional ability to act on his wants, to do otherwise *were he to want to do otherwise*, though in the present phi-fi case he lacks some weak conditional ability to alter his will, to will otherwise, due to Intervener. So, if he counterfactually wanted to vote Republican, he would have done so.

I had assumed this interpretation goes beyond what is necessary and might entail an unnecessarily-strong conditionalism. But Levin's argument comports with the weak-alternatives argument applied to CON: Though Agent cannot want otherwise in the actual world, he could want/do otherwise in the counterfactual world with different world-history/laws. Though determinism leaves no actual-world alternatives, it is consistent with counterfactual-world alternatives; so, though Intervener leaves no actual-world alternatives, his presence is consistent with counterfactual-world alternatives, though, in Frankfurt-cases, his presence bars *some* very nearby alternatives *ex hypothesi*. All we need to do is to factor whether in the nearest world in which there is no active-Intervener-interference-disposition and Agent wants to vote otherwise, he does so. There should be plenty of such worlds nearby for most agents in non-phi-fi worlds.

The counterfactual in which Smith wants to vote Republican is a back-tracker, in that the supposition that he wants to vote Republican implies that there is a possible world in which Intervener failed to function some time in the past. Surely we can imagine worlds in which Intervener does not function, unless Intervener is a Cartesian demon whose will is necessitated.[106] Absent a necessitated will that is fixed across all possible worlds, it begs the

---

[106] That an all-good God could act as a *manipulative* Intervener is self-contradictory on most interpretations of God, so even theists seem untouched by this option. The closest to this role that could be played by God is that of a CNC (Mele, 1995), by choosing to create that world in which he knows that circumstances He sets in motion will lead to Agent's freely making a specific choice, say, one that God wishes to be freely made. But this is not manipulative.

question to build necessity into the description of Intervener's will, to define away this coherent logical possibility.

We cannot get caught up in the phi-fi powers of "bugbears". Human Intervener cannot control all possible defeaters of his control over Smith. There are nearby possible worlds in which a black-out disables the implant, a meteor shower interferes with transmissions governing the remote control, or a heart attack that renders Intervener unconscious. Such contingencies permit Smith to form such an intention. So, Smith *can* do otherwise *if* he wants to. Levin's view may be fitted into my account. His point seems to hinge only on the sort of Frankfurt-case in which Intervener blocks the desire *to want* otherwise, but we need not engage in exegetical work on how to read this or that of Frankfurt's examples or on the many Frankfurt-cases that have arisen in the literature. Interveners may conceivably be situated at just about any point in the process leading to/from an action, from prenatal times to times after which the consequences of one's doings are beyond the sphere of one's control. Likewise, as in the "Mind argument", so named because so many versions of it have appeared in the journal by the same name (Mele, 1995), mechanical but non-agential versions of these Interveners may be substituted for each of them, as may be natural/biological processes, the succession of which substitutions is intended to tend toward slippery slopes of lost intuitions (*cf.* Mele, 1995).

In some cases, wants are blocked; in others, choices; in still others, actions; in some, self-formation is manipulated; and in others, the consequences of well-planned, well-meaning actions are manipulated. It depends. My analysis supports the necessary adjustments at any/all such Intervener apertures, and makes clear when we may and may not infer that Agent has Levin-style weak conditional ability even when Agent *seems to fail* PAPW2, thanks to Intervener.

PAPW focuses on *causal authorship*, a *metaphysical* notion of *agential-causative responsibility* for bringing about action, but may be used to account for *moral responsibility*:

PAPWR1: Agent is *weakly* morally responsible *iff* he satisfies PAPW1.
PAPWR2: Agent is *moderately* morally responsible *iff* he satisfies PAPW2.[107]

The differences between weak/moderate responsibility explain cases where necessary/sufficient conditions for autonomy diverge when PAPW1 is satisfied, but not PAPW2. Perfect autonomy and responsibility parallel Nozick's treatment of *perfect* value-tracking in moral judgment, analogous to perfect truth-tracking in epistemic judgment: Both extremes are unnecessary for minimal satisfaction in their domains.

PAPW1-satisfaction is sufficient for minimal autonomy; PAPW2-satisfaction is for moderate autonomy. PAPWR1-satisfaction is sufficient for minimal responsibility; PAPWR2-satisfaction is for moderate responsibility. Minimal autonomy is sufficient for minimal responsibility; moderate autonomy is for moderate responsibility.

The PAPW1 idea is that agents act *voluntarily* if they *would have still done as they did* had they been able to do otherwise. PAPW1 makes *principled* why Frankfurt-cases preserve the intuition that agents without alternatives remain responsible, contrary to PAP. Since Agent's actual-sequence behavior occurs without Intervener-interference, Frankfurt and others hold Agent responsible though he lacked alternatives, so PAP is false. If we *subtract* Intervener, nothing about Agent would change. If Intervener alters nothing, Agent is responsible though he lacked alternatives. This is "subtraction argument 1".

This argument is flawed. A Frankfurt-case only adds *possible* intervention, for *actual* intervention would defeat Agent responsibility outright. Next, it notes that intervention

---

[107] "Moderate" is left vague to fit the nature of the subject. "Strong" is reserved for indeterministic versions. F&R (1998) distinguish "weak", "moderate", and "strong" reason-responsive-mechanisms; they also cite vagueness as reflective of the subject.

wasn't actualized, that *if it was* it would *remove* Agent-responsibility, but that non-intervention is *non-removal* of Agent-responsibility. So, by introducing the possibility of a responsibility-defeating condition, then subtracting it, responsibility seems to remain.

When I was a child, my father would put up both hands and display 10 fingers, remove one finger at a time from one hand, saying "10, 9, 8, 7, 6" as each was removed, then say "6+5=11", where "6" was used to refer to the removed fingers, and "5" to refer to the fingers left visible. Subtraction argument 1 commits something like this "11 fingers" fallacy.

Nothing in Frankfurt-cases shows that there was non-zero Agent-responsibility *to begin with*, but only that if Intervener would have removed it in intervening, he leaves it untouched by not intervening. But this doesn't imply it was a *non-zero* amount. The Frankfurt argument thus commits a fallacy; just because *there would be no responsibility* if Intervener acted, it does not follow that *there is responsibility* if Intervener doesn't act. We cannot move from zero to non-zero responsibility by an 11 fingers move. Libertarians object that Intervener smuggles in determinist notions required to see a "sign" of movement toward the alternative, required to control Agent, which begs the question against indeterminism. But they can also object that determinism entails *zero* pre-Intervener responsibility, a view shared by hard determinists.[108]

But indeterminists insist that freedom requires alternatives, so Frankfurt-cases are just what can help decide whether alternatives are required for freedom and responsibility. Careful bookkeeping is needed about who is begging the question, etc., but incompatibilists cannot object to Frankfurt-cases on the ground that there is an antecedent *assumption* of zero responsibility per se, since soft determinists do not share that assumption. *Claiming* prior

---

[108] See, e.g., Ekstrom (2000, 2002) and Widerker (1995, 2002) for versions of these libertarian objections; see also Mele and Robb (1998) for a defense against the objections.

responsibility is one thing; using *11 fingers* to ground it is another. Opponents of PAP cannot just *assume* non-zero antecedent responsibility, left untouched by Intervener in subtraction argument 1. But they have offered only an *ipse dixit* that antecedent non-zero responsibility is *intuitive*. Allowing that Agent *would not be* responsible *if* he is Intervener's puppet does not entail he is responsible otherwise, *simpliciter*, or independent of the hypothetical. That's 11 fingers. Fischer not only commits such an *ipse dixit*, but actually defends the move.

> The first step is to argue [in Frankfurt-cases] that intuitively ... alternative possibilities are irrelevant to ascriptions of moral responsibility. One is supposed to see the irrelevance of alternative possibilities simply by reflecting on the examples. I do not know how to *prove* the irrelevance thesis, but I find it extremely plausible intuitively. When Louis Armstrong was asked for the definition of jazz, he allegedly said, "If you have to ask, you ain't never gonna know." I am inclined to say the same thing here: if you have to ask *how* the Frankfurt-type cases show the irrelevance of alternative possibilities to moral responsibility, "you ain't never gonna know." (2002, p. 292; emphases in original)

But Fischer loses his footing, for the mere presence of the non-zero intuition in the subtraction case does not imply that it holds in the non-subtraction case. The intuition needs justification. But the non-zero impression only seems intuitive relative to the 11-fingers subtraction maneuver.

That nobody intervened is *consistent* with Agent's responsibility, but consistency is insufficient here; 11-fingers cannot supply what's missing. Enter PAPW1, with its analogue of a *definition of jazz*. The idea in PAPW1 is that Agent does what he wants to do *whether or not he has alternatives*. He *usually* has alternatives – a *ceteris paribus* assumption for all intentional behavior – and selects from among them. One may object that he merely *believes* he has alternatives, but the claim may be made metaphysically: Intervener represents *the blocking of metaphysical alternatives*, since if any are entertained, he will block them.

PAPW1 factors for Agent's intention in the nearest world in which Intervener fails

and Agent *has* alternate opportunities or RFAs he *can* act on, to determine if Agent still acts on the same reason he acted on in the actual world. If he performs the same action, then the action based on that reason was not performed against his will; thus, it was sufficiently authored by him. If he *does* otherwise in the nearest world in which all his intentions remain fixed but Intervener is absent, then *this* "subtraction argument 2" indicates that his acting in Intervener's presence was influenced in a responsibility-undermining way by Intervener; thus, it was not sufficiently authored by Agent. PAPW1 thus renders subtraction argument 1's jazz effable, and explains why agents still seem responsible though lacking alternatives: They are acting on their own volitions.

If Intervener's *presence* implies *removal* of alternatives, then his *absence*, either *simpliciter* in non-phi-fi cases or *post-subtraction* implies *presence* of alternatives. This "subtraction argument 3" is a possibility-symmetry demand. Determinism entails there are no strong alternatives in Intervener's absence, only weak ones, so Intervener's presence cannot be what explains the absence of strong alternatives, for Agent is not exempt from determinism. Strong non-alternatives are *over-determined* by Intervener and determinism.

Either the absence of Intervener entails strong alternatives or not. If it does, Frankfurt-cases *support* PAP *in non-phi-fi cases*. If it doesn't, because determinism is true, then mere absence of Intervener *tout court*, *simpliciter* or post-subtraction-argument-1 does not guarantee the presence of strong alternatives, for determinism and Intervener simply overdetermine non-alternatives. Intervener's addition and then subtraction (11 fingers) has no effect on the underlying determinism-based non-alternatives. But if so, then anti-PAP proponents of subtraction argument 1 cannot assume non-zero responsibility to begin with, which is required for subtraction argument 1 to go through. That Agent is responsible independently of subtraction argument 1 needs support. But determinism and PAP are taken

to undermine that support.

Subtraction argument 1 is motivated by the worry that determinism removes alternatives, whose absence threatens responsibility, and aims to show that non-alternatives do *not* entail non-responsibility. *But that only works by antecedently assuming that agents are already responsible.* But if one thought agents already were responsible, one would not need subtraction argument 1. If one thinks subtraction argument 1 provides Fischer's jazz, one has lost count of a finger. But PAPW1 explains why agents can remain responsible absent strong alternatives.

I do not press the idea that requires equally-strong possibilities in Intervener's absence, since determinism entails absence of strong alternatives independently. I press the weaker-possibilities claim that in Intervener's absence the nearest world in which Agent is able to do otherwise *is closer* than the world located by moving outward from an actual world defined by the Intervener's presence. That is, in Intervener cases we must go much further out from the actual world to test PAPW1 than in cases of no Intervener.

In the nearest world where Agent would want otherwise, he is blocked by Intervener. *No Fixity violation is implied*, for we *already* entertain *that world* when we grasp the claim that Intervener would block Agent. We have to go some way out to a world in which Intervener is absent. In non-phi-fi cases, the nearest world to be entertained is one in which Agent wants to do otherwise, and in that world he can, *ceteris paribus*. PAPW1 only asks whether in the nearest world in which Agent can do otherwise, he *does* do otherwise or sticks with the same reason/action. Sober's weathervane metaphor (1995) suggests that he is *stuck*.

Levin might insist that since Intervener is supposed to block his *wanting* otherwise, he does not block Agent's *doing* otherwise. If he wants to do otherwise, hasn't Intervener already failed? Again, depending on where we stipulate Intervener's disposition of

intervention, we apply PAPW differently. Perhaps Locke's unwitting prisoner better illustrates how weak and moderate versions of PAPW coming apart, on Levin's interpretation of Intervener.

These differences matter more when we shift from *strong* to *weak* possibilities. Agent has no strong alternatives under determinism, irrespective of Intervener, but has weak alternatives, Bok's "practical possibilities" (1998), *which are, importantly, what he loses with Intervener.* Types of non-alternatives – necessities – cannot be homogenized, contrary to modal fallacy α.

Bok's practical/theoretical and our weak/strong distinctions overlap. For Bok, any option Agent considers *would* open a metaphysical path *were* Agent to choose it. So, from the vantage of Agent's power to open a path by selecting it, all such options are real possibilities, though only the selected one is metaphysically open, given determinism. It is analytic that I will do whatever I will do, from which it follows that I cannot *do* other than what I *ultimately* do. But this entails no loss of *control* over – ability to *select* from – available options. Each practical possibility in the vicinity of being chosen represents a path in a nearby world which would be open, were it endorsed. For it to be endorsed, Uniformity requires a different past, but this entails no inability.

Agent's weak/practical possibilities may be blocked by Intervener's preventing Agent's considering them. When Intervener is absent, weak/practical possibilities and related counterfactuals remain. The blocking of epistemically-accessible/entertainable practical possibilities causes the blocking of the related counterfactual possibilities Intervener controls, but these differ. We may distinguish practical possibilities *epistemically* or *metaphysically*. If Intervener permits practical possibilities *to appear in* Agent's

deliberations, *but only appear*, they are merely *epistemic*; if so, then they are not linked with the non-phi-fi-counterfactuals with which they are otherwise linked. Intervener allows the option to pop up in the subject's mind, but makes sure it is never chosen. Levin would agree that this is consistent with the following: If the subject chose the dangled alternative, he would have done it.

A tacit assumption is that Intervener permits epistemic possibilities to enter into Agent's deliberations, but stands by ready to sever their non-phi-fi-link with their counterfactuals, but this assumption is not innocuous. The epistemic/metaphysical ambiguity of Intervener's removal of alternatives has been overlooked, but is crucial, for Intervener cannot remove only metaphysical alternatives. I'll say why later, but if that is correct, it proves subtraction argument 1 is invalid. The loss of practical alternatives – as if one could only select from *one* alternative – bears on culpability in light of the Kantian point that motive/RFA is key in assessing behavior. Though strong/metaphysical possibilities are irrelevant to PAPW1-responsibility ascriptions, subtraction argument 1 doesn't establish this. Nor does it follow that weak/practical possibilities are irrelevant to responsibility ascriptions. For non-phi-fi Agents have practical alternatives – a background condition on all intentional behavior – and select from among them.

This reveals *highly-subtle temporal-indexing problems*, for in selecting "RFA_1", Agent either went through some process prior to Intervener-presence and had practical possibilities or he didn't. If not, then Intervener does reduce Agent's alternatives before Agent 'selects' RFA_1, which entails *Intervener contamination*, for there are *no other entertainable RFAs to select from.* If Agent did engage practical reasoning *prior to* Intervener-presence, then his having had *access to practical possibilities* and his having *selected from among them prior to Intervener's presence* is what accounts for the intuition

that he is *responsible*.

We may divide the deliberative process in three. The *formative* stage is when Agent entertains alternatives. The *selecting* stage is when Agent selects the RFA that will determine his action, Frankfurt's 'volition' and Davidson's 'primary RFA'. The *executive* stage is when Agent initiates the RFA-guided process that leads to action.

F&R (1998) treat my threefold process as two-fold. Since Agent can be receptive to a *best* reason, yet select a different one, *recognizing* involves *cognition/receptivity*, but *selecting* involves *conation/reactivity*. Reason-responsiveness involves both an *inner* mechanism of reason-receptivity, and *outer*-directed reason-reactivity, that involves bodily movement and the "outer" path leading from bodily movement, say, pulling a trigger, to an external event, say, shooting the Mayor (1998, p. 107). This division figures in responsibility for external events: Agents are responsible for the *consequences* of their inner-mechanism behaviors *iff* the outer path from inner-mechanism to event exhibits appropriate responsiveness to inner-mechanism (pp. 107 ff.). To trace outer events to inner-mechanisms, they use an analysis of causal/counterfactual conditions similar to ours. Since an RFA other than the-best-recognized-RFA may be selected, selectivity divides into *recognition* of best-RFA and *adoption* of primary-RFA.

With these divisions in mind, we must reconsider *the logic of apertures* for Intervener. Intervener cannot be around *in the formative stage*, or else he *must* remove all alternatives but one from the epistemic-menu/cognitive-horizon (Levin, 1979) of practical possibilities that shapes both receptivity/selectivity stages. This Intervener-contamination-suggestion runs afoul of the tacit assumption in Frankfurt-cases that Agent possesses a *spectator-like/epiphenomenal* detachment toward his options, present up to the choice-point. But deliberative phenomenology and physicalism suggest a deterministic-Kane-type process

where one is "of two minds" that *necessarily* engage *different psychoneural elements*, creating conflict in the mind-brain (2003). Agent *actively considers moving toward entertained alternatives*, but for each alternative other than Intervener-preferred-alternative, Intervener *must* intervene, *ex hypothesi*. If Intervener is around *only at the selectivity-stage/choice-moment*, he is functionally-equivalent to determinism, but only in the *que-sera-sera*-fatalistic sense that determines that Agent selects just the one thing which is *whatever* he will ultimately select. But this is not a control-undermining factor, and so even Intervener cannot undermine control on *this* Intervener-aperture option.

If Intervener is only around *just prior to the choice-moment*, and Agent inclines toward any alternatives as one might be nearly-viscerally-pulled to order shrimp right up to when he orders steak, Intervener *must* remove alternatives. We will discuss the contamination issue when we discuss the internal/subjectively-available character of RFAs in Agent's cognitive landscape.

Another assumption of Intervener-aperture-logic that needs review is that Intervener-presence implies he *controls* Agent, though he doesn't intervene. First, if the assumption *that control need not manifest as interference* is correct, the argument fails, for if control is present *whether or not* he intervenes, *then his presence cannot be subtracted*. For mere non-interference is what is considered the subtraction of his presence, but mere non-interference doesn't remove his control. Second, if the assumption *of non-intervening control* is valid, then whenever there is non-phi-fi-Agent possesses non-intervening counterfactual control over himself. This is "subtraction argument 4", and the sort of control it validates is "Intervener-style-self-control". And here's the dilemma: If Agent possesses Intervener-style-self-control antecedently under determinism, then there is no need for any anti-PAP subtraction argument 1. So, either subtraction argument 1 fails or is otiose.

Intervener-style-self-control is intuitive and doesn't need much support. For if *having the option of intervening but not exercising it unless such exercise is called for by behaviors deviating from the putative controller's plan* equates with *control*, then Agent possesses control when his action-sequences/willings conform to his action-plans/will-preferences. And he exercises control when his action-sequences/willings begin to deviate from his action-plans/will-preferences and he *steers them back into line with the plan/preference. A fortiori*, ubiquitous experience confirms this, and PAPW-satisfaction establishes it. This is autonomy.

Runner "R1" plans to run 5 miles, but exhaustion threatens to stop the run prematurely. Upon identifying these "signs" that controllee-part of R1 is about to deviate from controller-part's plan, controller-part exercises control by intensifying efforts against controllee-part, placing R1 safely back on course. (The metaphysics of intrapsychically-divided-self are divisible into hard/easy problems, but our Frankfurt-Perls-Kane model intuitively solves the easy problem of self.) Since controller wants to do other than what controllee is doing, so he changes his action in accordance with his desire to do otherwise, R1 satisfies PAPW2.

This constitutes *actual-sequence/actual-world* confirmation of Intervener-style-self-control, relative to R1's actually-manifest motives. Without actual-world manifestations of successful will contrasted against occasional failures and a background of successful intention/action sequencing under R1's control, mere counterfactual testing would lack inductive support. All of R1's will-activity provides the inductive data or reference base against which counterfactual judgments are more or less probable for R1.

PAPW2 takes into account that Agent more identifies with the fact that he wants to want to run, say, given his long-term commitment to marathon training, despite his not

wanting to run now, which Agent views as an alien or deviant want, so he puts himself back on course. The part of Agent that has a 2nd-order want is akin to Instructor who wants to regulate Driver who begins to drive the wrong way, whereas the part of Agent that wants not to run is akin to the driving student who starts to drive the wrong way; this is an intrapersonal-control analogue of an interpersonal-control scenario.

The counterfactual in PAPW2 here is only visible in the actual-sequence across the time-span during which he struggled with his desires at different orders. In other words, he has at T a strong 1st-order desire not to run, but also a momentarily-weaker 2nd-order desire to run, and the former is slowing him down; at T+n he has an effective 2nd-order desire and so he picks up the pace again. This is an actual-sequence modeling of what is normally seen only by comparing actual/counterfactual sequences: In the actual sequence, the appearance of deviant desires is like counterfactuals where Agent wants to and does otherwise; here, Agent wants to and does otherwise by picking up the pace. That is akin to Instructor's removing control from Driver. This is internal control of identified over alien desires.

To the extent Intervener-style-self-control is an ability to *redirect* Agent, it is superior to the guidance-control Fischer holds sufficient for responsibility (1987). Driver exhibits guidance-control when his muscular efforts and intentions control the direction in which the car moves; Instructor's regulative-control can override Driver's guidance-control. Fischer considers regulative-control the ability to *actually* do otherwise, thought to require Fixity-violating strong alternatives. But Intervener-style-self-control doesn't involve ability to do other than what one *ultimately* does, but, rather, an ability to alter what one is doing.[109] This

---

[109] *Not even God can do other than what they ultimately do*, for the reflexive attributive indexicality of "whatever happens" makes this analytic; erroneous equation of this tautology with Present-Fixity is analogous to erroneous equation of Past-Fixity with $\sim(P_0\&L)$.

intermediate Intervener-style-self-control is "pseudo-" or "weak" regulative-control, "ability to do otherwise" *naturalized*, "weak conditional ability", *not* ability to do other than what one ultimately will do.

If Driver is not *in contact with* cruise-control and other levers that guide the vehicle, he *lacks hands-on guidance-control*, but he exhibits *some* regulative-control *if it's true* that he can contact/manipulate the levers, should he so wish. This is counterfactual control; Intervener exhibits this. This control is proved whenever control-levers are contacted/manipulated in ways that fulfill controller-intended teleological patterns, but they also exist when not being exercised.

Fischer's *hands-off* control is regulative; he thinks determinism implies that since we can never do otherwise, we lack regulative control[110] – freedom. But this confuses *inability to do otherwise under different intentions*, "mundane conditional inability", and *que-sera-sera*-inability. We need not violate "*que-sera-sera*-Fixity" to possess weak regulative-control. Viewing inability to violate *que-sera-sera*-Fixity as a weakness is fatalism. The question is whether control requires access to divergent futures. Our analysis does not require this.

Hard determinists might demand that if Intervener did not intervene, intervening was not an option. One cannot say *both* "*if* Agent tries to X, then it's determined Intervener blocks X" *and* "*if* Agent doesn't try to X, then it's determined that Intervener doesn't block X", *nor also* "Intervener can block Agent" (*simpliciter*), not if there are no "if's" or "can's" *tout court*. '*Unexercised* control' begs the question whether 'could' is sensible when it is

---

[110] Animals possess mundane conditional ability (Shatz, 1986); they can act on variable intentions (Frankfurt, 1971). Since conditional ability *to do otherwise* is insufficient for *freedom*, Frankfurt proposes ability *to will otherwise*. But the will that inclines toward food is one the rat can restrain in light of the nearby cat. Most voluntary behavior involves mundane conditional ability, but not conditional ability to will otherwise. PAPW captures both.

determined that it not be exercised. In other words, Intervener is given powers the very advocates of the argument deny are possible. Given determinism, pessimists insist, intervening was never an option, so neither is Intervener's counterfactual/regulative-control over Agent. So, one cannot claim by a subtraction argument or symmetry thesis that Agent must equally possess regulative-control in Intervener's absence. "0x0=0", he may add.

Frankfurt-cases were supposed to constitute conditional proof against hard determinism: If we suppose a case without alternate possibilities, we can still have responsibility/freedom. What was supposed to make the non-alternatives plausible was a case where alternatives are absent even under soft determinist models involving agents with control being removed from them by other agents who can control them. But the details of the thought-experiments presuppose not that alternatives do not exist, but that they do – in fact, not only that they do, but that they constitute the ubiquitous background of agency against which they are artificially-removed by another agent who also has alternatives. But Fischer's "jazz" is only intuitive to the soft determinist. On this "hard-determined-Intervener-objection", whether he intervenes or not will not help the soft determinist: If he does, it was not up to him to do so, and so he cannot be held responsible for doing so; if he doesn't, he never had control anyway. So, the design of the thought-experiment is contaminated. I couldn't agree more: *Accounts resting on "jazz" collapse.*

Since the hard-determined-Intervener-objection is reasonable, such assumptions from Frankfurt-cases as premises for the conclusion that there is such a thing as Intervener-style-self-regulative-control would be unacceptable. But those are not premises in my account; my critiques of actualism, α, and β eliminated the grounds upon which pessimists may claim it is question-begging to presuppose *any* possibilities. *Weak* possibilities were not

presupposed, for Frankfurt-cases precisely *model* the blocking of *strong* possibilities. PAPW1 is designed precisely to capture the voluntary in the absence of alternatives *of any kind*. PAPW2 is designed to capture the causal efficacy of Agent-intention traditionally sought in conditional ability. Thus, weak-regulative-control is independent of assumptions from Frankfurt-cases, but was introduced therewith because intuitions are vivid for the Frankfurt-cases jazz-cognoscenti.

Weak regulative-control exhibits the dispositionality PAPW captures. The idea in PAPW2 is that when agents want to do otherwise, they do. R1's motivational system starts to direct the functioning of his whole system so that he wants to do otherwise than what he is beginning to do, which is to throw in the towel. Since he wants to do other than what he is actually doing, he does otherwise. If runner "R2" who quits after 3 miles, he still exhibits guidance-control, for his akratic intentions guide his quitting. But he fails to exhibit weak regulative-control, for although, like R1, he selected the 5-mile run from among entertained alternatives, he failed to carry that out. Both exhibit guidance; only R1 exhibits regulation.

There is a temporal factor involved here, as earlier. R1 and R2 both wanted to run 5 miles; only R1 succeeded. One might object that when they hit the 3-mile barrier at T+n, R2 does what he wants at T+n (throw in the towel), so it is not that R1 is doing what he wants and R2 is not, but that R1 is doing what he want*ed* at T, and R2 is not. But R2 is not doing what he *overall wants* during span-S from T to T+n, for the better part of R2's motivational system is committed during span-S to the 5-mile run, even at T+n. R2 fails PAPW2 with respect to long-term/span-S-committed desires, but passes PAPW1 with his short-term akratic desire.

Intuitively, akratic behavior is loss of regulative but not guidance-control. That agents regret akratic behavior coheres well with the idea that they normally exercise weak

regulative-control over activities of similar sorts, *without which background of typical control it would not make sense to adopt those reactive attitudes towards oneself* on the mere ground of guidance-control over these akratic activities.[111] This idea of *a background assumption of autonomy* needs to be incorporated into the metacausal theory as the *agential functionality norm* against which individual cases must be judged. We do not circularly assume a background of autonomy to explain autonomy, for the dialectic was not *first* establishing autonomy by reference to the background assumption; other grounds were offered to think we have a background of ubiquitous autonomy. The claim here is that *this background* needs to be referenced to sort out intuitions in individual cases. The general background of autonomy is an essential feature of the metacausal conception: Autonomous agents are generally autonomous in most of the domains of their intentional behavior, and it is by reference to their unique constellation of autonomous functional specifications as idiosyncratic agents that test cases need to be adjudicated.

Weak regulative-control differs from regulative-control, which can only be possessed by agents in indeterministic worlds with access to strong alternatives. Weak regulative-control, however, can be possessed in deterministic worlds in which agents only have *practical* access to weak/counterfactual possibilities, but no access to strong/metaphysical possibilities. It is enough if Agent accesses *practical* possibilities, and acts on them. Doing so determines his future, for on determinism everything in nature has causal powers, and what distinguishes determinism from fatalism is precisely this intuition – that our deliberative behavior is what determines our futures.

Recall Fischer's similar point about *reason-responsiveness* (1987). Agent need not

---

[111] Not all reactive attitudes are sensible: Studies show women feel responsible for breast cancer, though they know it doesn't result from lifestyle choices (Benson, 1996, p. 264).

access an alternate world by *doing* otherwise to exhibit guidance control, so long as his deliberative-mechanism is reason-responsive in the actual-world sequence. To see if it is, we *imagine* varying the circumstances – we don't *actually* vary them – say, a stronger reason that should defeat Agent's choice. If it does, Agent's deliberative-mechanism is reason-responsive; if not, it is not reason-responsive. Counterfactual testing doesn't require Agent-access to alternate worlds, but just the judgment whether the mechanism would/wouldn't respond differently under the influence of better/worse reasons. The truth-conditions are *externalist* and have nothing to do with Agent's entertaining of alternatives; *our justification* that a real Agent satisfies such conditions can only be made inductively, based on our knowledge of his general background of agential effectiveness.

This defense applies exactly equally to both models, Fischer's and ours. However, the shift from *Agent's* counterfactual abilities to the *reason-responsive-mechanism's* counterfactual abilities packages counterfactual abilities just covertly enough to slip by the hard determinist. But the shift to mechanism reduces agency to something on par with a thermostat. A climate-responsive-thermostat exhibits climate receptivity (it properly detects the temperature) and climate reactivity (it properly adjusts the temperature to desired settings) in the same way a reason-responsive-mechanism exhibits reason-receptivity and reason-reactivity. This is not a bad thing, per se, but its focus on *mechanism* misses something: agency.

PAPW helps distinguish actions that *manifest* in full accordance with Agent's will from those that *originate* within Agent, with or without full consent or control. Libertarians and other ultimatists want endogenous origination, but exogenous origination does not pose a problem if Agent maintains Intervener-style-self-control over intentions. For Frankfurt (1971), origination in Agent is inessential if Agent approves of the intentions, and on this

ground he eschews a causal control analysis. But a more robust ability is one that can also disapprove *effectively*, which amounts to metacausal control.

All that weak ability (PAPW2) entails is that Agent would have done otherwise under a different motivational history, i.e., if he wanted to. That it is determined that he *does not* want to do otherwise doesn't entail that he *cannot* want to do otherwise, as actualism supposes, but only that he *doesn't* so want. (If a certain phi-fi-Frankfurt-case is an exception, that proves that *it is not analogous to determinism*, which doesn't entail actualism.) For R1 evinces, there are countless actual-world cases in which we want and do otherwise. The inclination to deny that there are actual-world, mundane-conditional-ability cases of wanting/doing otherwise is a CON holdover. For Agent's PAPW2-satisfaction implies Agent's wants are effective without strong alternatives. Thus, it makes sense to speak in the ordinary way – that we actually *do* otherwise whenever we change our minds, form new intentions, or battle old ones. This is autonomy.

J need not violate Present-Fixity or render P false, since J's ability is never manifest as a Fixity-violating *change* in actual-world history. Since J need not violate Fixity or render P false, J need not render false P_0&L either. J exhibits autonomy within his Fixity-compliant world when he satisfies PAPW2, evincing weak regulative-control over his actual-sequence actions and thus over *his* segment of history. Determinism implies autonomy *must be* actual-sequence type, or else it won't be Fixity-compliant. PAPW2 satisfies such demands.

J's power manifests differently in counterfactual worlds in which antecedent causal conditions (C_J) are different, producing different actions. When counterfactual-conditions C_{J1}, C_{J2}, ... C_{Jn} hold, respectively, state descriptions referring to J's different actions (at T) P_1, P_2, ... P_n result, accordingly. This principle is inductively-confirmed *within* the actual world by countless observations where relevantly-different intentions of agents appearing to satisfy

PAPW2 standardly result in concordantly different actions. Again, this is mundane conditional ability.

No Fixity-violating abilities are needed to satisfy PAPW2, only the everyday ability to do different things when we want to. When we want to do different things and do them, PAPW2 is massively confirmed, but when we want to do otherwise but fail (*akrasia*), we only satisfy PAPW1; when we chronically fail, we may also fail to satisfy PAPW1 thereby. In these cases, PAPW2 is violated, though PAPW1 may be satisfied, but its violation explains privation cases.

Akratic behavior is *local* failure of PAPW2, since it is typically only *occasional*, and typically involves only a small *subset* of motive/action types. Compulsion involves *global* failure of PAPW2 because it is *chronic*, though it does not usually involve the *entire* set of motive/action types. If *all* of X's motive/action pairs chronically failed to satisfy PAPW2, it is doubtful we would be inclined to consider X sane, a moral agent, or a person.

Frankfurt's "wantons" are chronically deficient *locally*, relative to a *subset* of motives, but that is too weak a ground for exclusion from personhood, though perhaps it applies to those deficient both chronically/globally.[112] But personhood need not be denied to those who suffer chronic *privation* globally, relative to the *entire set* of motives. For in order for it to *be privation*, they must have previously enjoyed the use of this faculty. Illness can place a person in this category *temporarily*, and they can experience it *chronically*, but still recover. Intuitively, they remain persons during privation, as dreamers remain each night. So, if they injure someone or run up bills during lapses, they are still partly liable for their

---

[112] Frankfurt (1971) restricts "wanton" to *non-privation* cases – those *incapable of failing* to carry out their will with respect to desires of type d because they lack a hierarchical d-division and so cannot suffer *privation* of d-range autonomy. While many competent adults are wantons in this (local) sense, they are not what philosophers (globally) mean by 'nonpersons'.

actions. The background assumption of – resilient, if minimal – personhood maintains liability, and its abeyance is what accounts for mitigating/exculpating intuitions.

Both occasional/chronic failure are typically local. *Sufficiently autonomous agents* may instantiate both. For they maintain a *background* of enough PAPW-satisfying activities – e.g., crossing the street when they want to – to count as autonomous. Agent must attain global functioning as a person to suffer its privation. Consider catatonia and Parkinson's disease.[113] There are some forms of schizophrenic and psychotic behavior that exhibit more global forms of PAPW-violation, e.g., motor disturbances. It makes sense to view them as insufficiently autonomous, thus not moral agents, even if still – personhood-challenged – 'persons'. So, along the spectrum between cases where PAPW is/isn't satisfied, our judgments about personhood, responsibility, and exculpation fit nicely.

PAPW tests close-range (Russell, 2002) behavior: whether *volitions* produce *behavior*. The ubiquity of apt linkages between volitions/behavior constitutes *massive confirmation* for daily autonomy *qua* PAPW-satisfying control of volitions over actions, and of their power as variables: When others are present, we do other things. The justification for weak/mundane conditional ability is in numerous counterfactuals, confirmed in countless generalizations about daily experience. Thus, when we talk about Agent's weak conditional ability, we do *not* mean he has *access* to world W_2, such that in choosing he *determines* which world he *makes* actual. He comes close, however, for his selection from among entertained possibilities does determine the subsequent state of the actual world. But his selection is not from an item that would connect him with another world or violate Fixity. His selection is necessarily from the world he occupies; of the unselected items he entertains,

---

[113] Both suffer akinesia, but the broken links in their intention-action sequences differ (Northoff, 1999).

all we mean is that *had he wanted otherwise*, which would put his past in W_2, he'd have acted differently. That translates into his wants being effective. Nobody violates *que-sera-sera*-Fixity, but what they ultimately do is a function of their control over their wills and of their wills over their actions.

CON conflates *que-sera-sera*-inability with conditional inability. If ultimate non-control over choice renders Agent's intentions causally impotent, then mundane conditional agency is a global illusion. There is little evidence to support an error theory about what amounts to massively-confirmed experience. Studies like Milgram's (1963) (test subjects perform cruel actions under the assumption of experimenter authority), Nisbett and Ross (1980) (shoppers unconsciously select the right-most item and invent reasons afterward), and the like provide at most the basis for what on analysis amounts to no more than a hasty generalization. These cases, like unconscious Freudian motives, are not representative.

We are not globally deceived about the causal efficacy of our wills. The global illusion claim equates us all with intentional zombies that happen to have a spoonful of epiphenomenal consciousness, a sprinkling of catatonia, and a wallop of self-delusional beliefs and impotent desires that all coincidentally map the contents of their epiphenomenal mental states onto their actions, making each of them appear intentional, willful, coordinated with their efforts, etc.

Odd bed-fellows, Descartes and Wittgenstein can help here. The arguments Descartes gives toward the end of the *Meditations* that rest on the *coherence* of the evidence in favor of the waking state (a thread of continuity in space and time, coherence, etc.) intuitively support the claim that such a complex illusion is incredible. Autonomy equally enjoys massively-coherentist confirmation, but illusionism has little to recommend it; the evidence strongly supports autonomy, but largely *underdetermines* illusionism.

Wittgenstein's arguments against private language apply here. Without confirmation from a common public reality, our level of complex agreement in experience would be impossible. The public world of agential action is a necessary condition for the agential way of life for each individual among billions who coordinates his agential, reactive-attitude-rich life with others. Strawson makes a similar point (1962).

Assuming illusionism displaced, let us return to the issue of how our control is established, and compare our account with F&R's (1998). In ordinary agency, then, F&R say, guidance-control is exhibited in the actual sequence *iff* Agent's actual-sequence action-producing mechanism is *moderately reason-responsive*. To see if his mechanism is *weakly* reason-responsive, we need to see if there is a world in which his action-producing mechanism, presented with a different RFA, responds differently. If there is such a world, then the action-producing mechanism is minimally responsive to reasons, and exhibits *weak* guidance-control.

Since Intervener blocks Agent from doing otherwise in the alternate sequence, however, Agent cannot do otherwise, and so lacks regulative-control. If Agent possessed regulative-control, then Agent would possess strong reason-responsiveness. Determinism is equivalent, F&R aver, to Intervener with respect to blocking Agent from possession of regulative-control, for under Present-Fixity, Agent is only capable of acting at T in accordance with whatever singular outcome the past/laws determine for Agent at T. Determinism implies that there are never alternatives for Agent's actions at any T, so it rules out regulative-control. PAP requires regulative-control, but is unsatisfiable. But responsibility does not require regulative-control, as Frankfurt-cases are designed to show. F&R contend all that is required to explain the intuition that responsibility remains in the absence of alternatives is guidance-control, which is consistent with determinism, since it

involves only actual-sequence reason-responsiveness.

So far we have given only the conditions for *weak* guidance-control, i.e., there must be a world in which Agent's action-producing mechanism is responsive to a different incentive. But since weak reason-responsiveness could obtain under bizarre circumstances, F&R note, weak reason-responsiveness is insufficient for responsibility.

Suppose Jane wants to see the Rolling Stones perform live, but was only able to purchase a ticket for $1000, and declines. But suppose she will purchase one for $500. At first, she appears to be weakly reason-responsive, since there is a world in which she responds to different incentives differently. But now suppose she is offered a ticket for $450, but won't purchase it. Maybe she has numerological/astrological beliefs about details about the band with which she has become obsessed. In that case, her ability to respond to *some* reason(s) is insufficient for responsibility. More is needed, and to meet this need F&R suggest *moderate* reason-responsiveness, which obtains when one's reason-responsiveness under counterfactuals displays an appropriate *pattern* of recognizable rationality, which they define as a fit with socially-recognized standards of reasonableness in the relevant domain.

A mere pattern would not suffice, e.g., ticket prices formed by prime numbers or astrological patterns, which is why F&R tie the requirement to wider standards of rationality in Agent's community. Communitarianism only wards off counterexamples that target the over-inclusiveness of the pattern criterion, but resilient objections could be made by reference to doomsday cults and other irrational communities. This is a problem, but there is a bigger one.

By switching from *Agent's* abilities to Agent's *mechanism's* abilities, F&R seek to circumvent the intuition that guidance-control indirectly relies on alternate worlds, which are ruled out by both Intervener and determinism. By questioning whether the *mechanism* has

the counterfactual disposition to respond a certain way in alternate worlds, while maintaining that *Agent* has no access to such worlds, F&R deflect – like bullfighters waving a red flag – the charge that guidance-control invokes alternate possibilities, and so violates determinism.

For *mechanisms* are the sorts of things it is intuitive to view as behaving in certain ways under certain conditions, and this way of thinking seems innocuous. But F&R buy into actualism, rule out regulative-control, and so adduce an actual-sequence account of responsibility that doesn't require alternate possibilities. But the very *notion* of a 'mechanism' makes no *actualistic* sense if it must have *dispositional* properties that manifest in alternate possible worlds. Thus, F&R smuggle in non-actualism via the seemingly-innocuous substitution of Agent's *mechanism*'s ability for *Agent*'s ability. The shift from *Agent's* ability to *mechanism's* ability cannot circumvent the smuggling charge, for if the *mechanism has* the dispositional ability, and *Agent has* the mechanism, then, by the transitivity of "has", *Agent has* the ability to respond differently under different circumstances – the ability to do otherwise. *Any actualism that denies Agent ability must also deny mechanism ability.*

F&R have done no work toward dispelling the actualist implications of CON. Rather, they embrace CON,[114] which is what gives rise to their attempted semi-compatibilism. Thus, they are not entitled to models that violate the actualism that motivates their semi-compatibilism, but the mechanism model implies *dispositional non-actualism.* Though they come close to laying the foundations for a refutation of actualism in their move to the effect that *Agent* need not access counterfactual worlds, they *just barely miss it* when they say that *the mechanism* has the dispositional properties, tested for in counterfactual

---

[114] Fischer has an in-depth analysis in favor of CON (1994).

worlds. The slightest adjustment to their formulae – plus lots of conceptual work – would have given them my anti-actualist account: The same tests that ground responsibility on their account ground autonomy on mine.

My account can make the same moves, but I have the dialectical luxury of not needing to hide dispositionality in Trojan-horse-type mechanisms: I can attribute this power directly to Agent without compunction. The metacausal features of my account permit me to defend a *robust, non-semi-compatibilism* in which Agent standardly has Fixity-compliant weak regulative-control, weak conditional ability. Recall F&R's Agent does not access other worlds, but to test whether his actual-sequence mechanism is reason-responsive, counterfactually possible worlds are considered to see if, under alternate scenarios, Agent's mechanism responds appropriately to alternate stimuli. A similar approach is available to test whether Agent satisfies PAPW.

Thus, if Agent satisfies PAPW1, his actual-sequence behavior exhibits guidance-control. This is visible in the nearest world in which Agent *can* do otherwise under otherwise-same conditions, but does not. If varying *only* the opportunities results in a different action in the counterfactual sequence, *lack* of opportunity may be *inferred* as a factor that influenced Agent's actual-sequence behavior, in which case Agent fails PAPW1, his behavior is not fully voluntary, and so *he lacks full guidance-control* over his action.

Similarly, if Agent satisfies PAPW2, his counterfactual-sequence behavior exhibits determinism-friendly regulative-control stemming from his will. This is, similarly, determined by a third-personal view of the nearest world in which Agent can do otherwise and wants to. If varying Agent's intentions immediately results in a different action in the counterfactual sequence, then Agent minimally satisfies PAPW2, and so possesses minimal regulative-control; further, we may infer that the guidance-controlling intention in the actual

sequence was not a factor that influenced Agent's actual-sequence behavior. In the counterfactual sequence, Agent responds *equally* to an *alternate* intention.

Were Agent to continually pursue intention X even under the presence of more attractive intentions Y and Z, this would reveal a problem with Agent's ability to act on X or not, and thus shows that X functions in the actual sequence as a constraining influence: If all other things are equal, but Agent's having another intention does not result in another behavior in the counterfactual sequence, then he does not satisfy PAPW2, he lacks regulative-control, and cannot do otherwise. This is why conditional ability (PAPW2) and not merely voluntariness (PAPW1) is necessary for a robust sense of freedom, contrary to the Frankfurt-cases jazz-cognoscenti. This strongly suggests that Frankfurt-cases do not support anti-PAP intuitions. Here, Agent must do X, but the type of modal force is not the mere "must" that attends all determined actions, nor even the stronger "must" that attended Luther's famous choice, but the "must" associated with loss of autonomy, e.g., in addiction, coercion, etc. That these different modal forces all may be referred to by the homonym "must" is no reason to homogenize necessities, contra α.

The presence of the guidance-controlling-intention ought not be a constraining factor that overly-influences Agent's actual-sequence behavior. A weak case of guidance-control that fails to satisfy this requirement is a counterexample to F&R's equation of guidance-control with responsibility. A super-resilient intention, like a targeting-mechanism in heat-seeking missiles that cannot be deflected, could satisfy moderate reason-responsiveness, yet fail to be under Agent's control in a sense sufficient to ground responsibility. Some suicide bombers may possess super-intentions that are reason-responsive relative to their otherwise-rational communities. We mustn't conflate *their being responsible* with F&R's explication

of *why they are responsible.*

Here is a valid horizon-worry about our lives being beyond our control. Though invalid in CON, horizon-worries may be valid in *individual* cases, where special features of Agent's early history overdetermine his subsequent character. It is a matter of *bad moral luck* if Agent is born into a fanatical fundamentalist community, say, raised in a Taliban grammar school, since those circumstances create a suicide bomber. Not everyone raised in certain circumstances is destined to become evil, but I am only isolating intuitive such cases.

Suppose "Suicide Bomber" exhibits weak guidance-control and maybe even weak regulative-control, in that if he wanted not to bomb the Israeli school bus under some scenarios, he wouldn't. The problem is that in most nearby worlds, Suicide Bomber still wants to bomb the school bus. We might have to go back to a pre-agential point in Suicide Bomber's life and fork from there into possible life-paths that radically diverge from his actual one to locate any world in which he does not want to bomb the bus. That he does not bomb the bus in some distant world does not show that he has sufficient regulative-control over that desire in the actual world. *The great distance between the actual and possible worlds* needed to assess these principles matters, though only an informal account of the role of distance is needed to support these intuitions.

If the fork needed to locate a world where he wants not to bomb the bus is *pre-agential*, that doesn't support Suicide Bomber's regulative-control *now*, though there may be a *very weak* sense in which he possesses *some* regulative-control, analogous to being *minimally* responsive to one *bizarre* reason. That he possesses *this* control is insufficient for responsibility.

If so, then F&R-style guidance-control is insufficient for responsibility in certain cases, though not in most others. What is needed is supplied in PAPM, which adds a

metacausal criterion to PAPW to yield more robust control conditions. PAPM is thus more able to sort out cases where necessary/sufficient conditions for autonomy/responsibility come apart. The informal idea in PAPM is that autonomous agents possess robust metacausal control over their actual-/counterfactual-sequence behaviors to the extent that they satisfy PAPW1/PAPW2 and *normally* maintain meta-conscious control over their doing so. PAPM is also applied to the formative years of agent-self-cultivation, and to every part of the action-producing system, any and every part over which Agent may have metacausal control. Agent has metacausal control over himself *iff* he has *the same sort of access to his own mental states and thus the ability to alter them that Intervener has over Agent.*

Thus, Suicide Bomber may not possess self-formative control, but he does have *current* control sufficient for responsibility, though he lacks moral luck and may have "blind spots" in his moral-reason-responsiveness, due to fundamentalist overdetermination. His lacking moral luck may generate *some* sympathy *for his childhood*, but not enough. Despite the overdetermination of beliefs in the vicinity of his terrorist behavior, he has enough mundane-autonomy resources *in the background of his daily life* to warrant the claim that he has the capacity to know better and to do otherwise, and his bad luck kicks in as misfortunate failure to turn this capacity into an ability.

On F&R's model, guidance-control *should* explain the Suicide Bomber case, but doesn't. The metacausal theory does: *The regularly-exercised capacity for weak regulative-control in the non-blind-spot areas of his life* reveals *a background of general autonomy and thus moral agency*. Compare Frankfurt's local, say, desire d-range wantons: In non-d areas of their lives they are not wantons, thus they are still persons. F&R cannot use weak regulative-control, so their account cannot explain why we hold Suicide Bomber responsible.

Let us turn now to the key principle of the metacausal theory, the *metacausal*

principle of alternate possibilities.

PAPM:	Agent is robustly metacausally self-guiding, self-regulating, autonomous *iff* he satisfies both PAPM1 and PAPM2.
PAPM1:	If Agent satisfies PAPW1 and has metacognitively-conscious guidance-control over the actual-sequence functioning of his action-producing system, then he exhibits *minimal* metacausal self-guidance and autonomy.
PAPM2:	If Agent satisfies PAPW2 and has metacognitively-conscious determinism-friendly regulative-control over the counterfactual-sequence functioning of his action-producing system, then he exhibits *moderate* metacausal self-regulation and autonomy.

"Metacognitively-conscious control" needs to be spelled out, for metacognition doesn't guarantee control, as seen in cases like Hamlet and Woody Allen.

Intervener has a form of *inclusively*-metacognitive control over Agent, since he has *cognitive* access to Agent's *cognitive* states, thus generating a 2nd-order cognition, though not within 'the same mind'. One might argue that if a second mind-body system is under one's control, such that when controller exercises an intention and controlled body executes it, controlled mind-body *just is* the controller's. One might object that my control of a stolen car doesn't make it mine, but things are different with mind-body systems, and here there are two senses of 'mine', moral and metaphysical.

If I control a stolen-body the way I control my birth-body, it is *mine* metaphysically, but not morally. If temporary, as in hypnosis, controller is responsible for the action but doesn't own controllee-body; controllee functions as a temporary extension of controller's will. If the take-over is permanent and involves the cessation of the previous psychological entity, then such a take-over would be morally equivalent to murder, though the second mind-body system remains homoeostatically-operative – no *habeas corpus*. The resultant mind-body would then be one's own, metaphysically. In the case of *total* Intervener take-over, Intervener has metacognitive control over second-body the way he has control over

another person, but here over his own 'second-self', since he then *is* both bodies, a scattered individual. Agents typically have this sort of Intervener-style-self-control, for these are, on reflection, equivalent.

Note that metacognition made control possible here, suggesting the principle discussed in connection with seizures: *Any increase in metacognition makes an increase in metacausal control possible.* If "possible" is removed from the last sentence, the principle is subject to many counterexamples. The idea is that looped consciousness of X makes its control *possible*. This is an extension of the generalization about feedback being a condition for the *possibility* of control.

Feedback about environmental orientation concurrent with proprioception of motor exertions is necessary for the sort of coordinated motor control we exert in daily movements. This is a basic organismic fact. Organisms operating solely on the unmediated-S/R-level possess the same elementary capacities for motor control as we do,[115] but lack the options that come with our metacognitively-looped meta-consciousness, which permit us to mediate between S/R. Motor control with looped consciousness is at the root of autonomy. Though it *is enough* for it to become a *capacity* for control, more is needed for the capacity to become an *ability*. For this, one condition is repeated *practice*, what converts mere bike-riding capacity into bike-riding ability.

The sort of control Intervener has over Agent via access to Agent's mental states is available to Agent in Intervener's absence, via Agent's access to his own mental states. Agents have these meta-abilities contingently and as a matter of degree. Those that *do* satisfy PAPM. The "action-producing system" includes all the input/operations/output

---

[115] On motor control and feedback, see Schwartz and Shapiro (1976) and Langer (1967).

stages discussed earlier when comparing F&R's bifurcation with my threefold distinction. To the extent Agent possesses metacausal control over any segment of his action-producing system, he possesses autonomy. For every division in the system, Intervener can find an aperture. Similarly, for every element of the system, some Agent will possess control over that element and others will lack control over it.

In light of the heterogeneity, ubiquity, and idiosyncratic nature of voluntary experience, the idiolectical hierarchical structure of the self-system, plasticity, and multiple realization, it is likely that the neurosignature-dependent causal-functional structures that embody the wills of beings with autonomy differ radically. This is why the autonomous will must be described by abstract causal/functional principles, with many qualifications. The vagueness in our theory should match that in the data. Idealizations set upper/lower limits: Any being who has 100% control over 100% of its behavioral system has *maximal* autonomy, and one with non-zero control over any element in its behavioral system has minimal autonomy. Yes, it follows that thermostats and snails have non-zero control over their behavior, which behaviors differ. Of course, thermostats are not conscious, so they lack what meta-consciousness makes possible for us. So?

F&R's guidance-control does not require access to possible worlds. Similarly, we only *consider* counterfactual worlds to test whether Agent has dispositional agential functioning. This disposition is *minimally* instantiated when PAPW1 is satisfied, and increasingly so when PAPW2, PAPM1 and PAPM2 are satisfied: when, in the nearest world(s) in which Agent can do otherwise, he does the same thing (PAPW1); when, in the nearest world(s) in which Agent wants to do otherwise, he does otherwise (PAPW2); and when his doing so in the actual (PAPM1) and counterfactual world(s) (PAPM2) is under his metacognitively-conscious control.

Talk of possible worlds here is externalist, or else Agent would be akin to epistemic candidates incapable of satisfying internalist conditions for knowledge. But just as a *knower* need not be able to prove that he is not a brain in a vat for *us* to be confident that he satisfies externalist conditions for knowledge, so, too, *Agent* need not be aware that he is not under the hidden control of Intervener for *us* to be confident that he satisfies the externalist condition for autonomy, PAPM. Nor need *Agent* perform PAPM tests for *his* beliefs about his autonomy to be grounded: The ubiquity/heterogeneity of his and others' wills in daily experience and the countless causal-counterfactual inferences drawn therefrom provide massive tacit confirmation. Though Agent need not *know* that he satisfies PAPM, *if he does* it is likely he reliably believes he is autonomous.

Though internalism is not apt for the specification of autonomy conditions, the RFAs Agent *considers* form his cognitive horizon in deliberation, *are* internal, and *do* shape his selection of RFAs that bring about his actions. *Internal* reasons play a *causal* role in deliberation, *not external* reasons of which Agent is unaware. Though Agent's receptivity to *external* reasons factors into assessing his *responsiveness* to reasons and thus his *responsibility*, his ignorance of external reasons only bears *indirectly* on the determination of his *autonomy*. If his disposition to recognize external reasons is of low caliber, this reduces the range of options that become his internal, subjectively-available reasons. Thus, the counterfactual worlds we need to view when assessing autonomy are determined by Agent's cognitive horizon, which demands an internalist view of his reasons. But this does not suggest internalism *simpliciter*.

Internalism is required for RFAs that determine Agent choice and affect which worlds enter the equation when assessing whether he satisfies PAPM. But PAPM-testing requires externalism. Internalism about RFAs invokes Bok's *practical possibilities*, the

RFAs *Agent considers* and from which he selects. The causal/counterfactual properties of these practical possibilities *within* Agent's cognitive horizon determine what Agent chooses. Since RFAs are causally effective, not epiphenomenal, freedom is causally grounded, naturalized, determinism-friendly. Bok's practical reasons, then, are not "merely epistemic", which is the only non-actualist impression that might seem to justify dismissing them.

Wolf (1990) argues that there is an asymmetry in responsibility. Agents are praiseworthy when they do what is right, but never blameworthy. She cites Luther's famous case and that of a woman who, upon seeing a child drowning, feels compelled to save him, simply because she so identifies with the child and is so attuned to the 'Good' that she cannot consider not saving the child. Agents seem compelled here to do what is right for good reasons. Wolf says they cannot do otherwise, given their virtuous natures, but are still praiseworthy. The criminal or other wrongdoer, however, also cannot do otherwise, given his vicious character, but it would be unfair to hold him blameworthy, since his turning out to have his character was not up to him, and his behavior flows from his character in ways in which he cannot control. The implicit justification for this disparate treatment of equivalent cases must be that *while it does no harm to praise the virtuous, it does harm to blame the vicious*. For though neither *deserve* our praise or blame, it would be acceptable to praise the virtuous, but unfair to blame the vicious.

F&R object to the idea that reason-responsiveness entails moral asymmetry (1998). The example they use involves auto mechanic Leno and his neuroscientist son Nick (1998, p. 59). Leno is plotting on his own to charge Mrs. Pratt for imaginary repairs on her Mercedes. Nick has planted a chip in Leno's brain designed to guarantee that Leno rips Mrs. Pratt off, *iff* Leno should show signs of weakening his intentions. Leno's intentions remain in force on their own. F&R dub this "Mechanic", and argue that since Nick's chip played no role, Leno

is blameworthy, though he lacked alternatives. So, the asymmetry thesis is false. Their explanation is that Wolf's examples are not representative. But comparison with F&R's own case of "Hero", one page prior, undermines this conclusion. In Hero, Matthew walks along the beach, sees a drowning child, deliberates momentarily, and jumps in and saves the child.

> We can imagine that Matthew does not give any thought to not trying to rescue the child, but if he had ..., he would have been overwhelmed by literally irresistible guilt feelings that would have caused him to jump into the water and save the child anyway.... Here is a case in which no responsibility-undermining factor operates in the actual sequence and thus Matthew is morally responsible for what he does. (*Id.*)

Because his alternate-sequence irresistible guilt didn't enter into the actual-sequence operation of his reason-responsive-mechanism, he remains praiseworthy. F&R conclude that "[t]aken together, 'Hero' and 'Mechanic' ... illustrate that Wolf's asymmetry thesis ... is false" (1998, p. 59). I agree that the asymmetry thesis is false. But their analysis doesn't establish this.

There are two types of inability to do otherwise in the cases of Hero and Mechanic. Matthew's is *internal*, and has to do with Matthew's *own* irresistible guilt feelings, whereas the Leno's is *external*, and has to do with an *external agent*, Intervener Nick. Matthew's own necessitating conditions operate in the alternate scenario, in which case he fails to satisfy PAPW2. (Levin would insist that there be a world in which he is able to want to not save the child.) But we are still inclined to praise Matthew because his own necessitating condition in the alternate scenario is a sign of a virtuously-cultivated character. Thus, though he lacks alternatives *now*, his autonomy may be "traced" back, using *their* tracing criterion, to his satisfying PAPW2 when forming that strong conscience, even if it is so strong now that he cannot but act on it. Thus, though Matthew directly fails to satisfy PAPW2 and apparently lacks alternatives in the present, he derivatively satisfies PAPW2, so Hero and Mechanic are asymmetrical.

I have avoided the dubious Aristotelian move of saying that we can always trace the life story of an autonomous person back to a choice, so that a person now is free because of what he freely chose when he was three. If the needed back-tracking passes beyond the pre-agential barrier, we part ways with Aristotle, and say Agent is not responsible in such a case.

Matthew's self-cultivation supplies the alternative-eliminating conditions, but Nick's chip in Leno provides the alternative-eliminating condition, so his condition cannot be traced back to earlier PAPW2-satisfaction. Thus, Matthew is responsible, though he lacks alternatives, because *he designed himself* so that he must do the right thing. This may smuggling in the warrant of responsibility from traced-back, thus derivative, conditions of responsibility, in which case we don't have a case of full responsibility without – earlier, but relevant – alternatives.

To test for symmetry between Hero and Mechanic, we need to see the difference in distance between the actual and the first possible worlds for Matthew and Leno in which they (a) are able to do otherwise, and (b) want to do otherwise. So, to test PAPW1, we need to go to the nearest worlds in which they are each able to do otherwise, and see if they still do what they did in the actual sequence. It matters how far away from the actual world such worlds are for each of them, to see how committed they each are to what they did.

Nick interferes with our calculations, for it is not clear how we are to determine what worlds to view to run the PAPW experiments on Leno. To test Leno's normal, non-Intervener-involving ability, which he lacks under the chip, we need to (a) go far out from the actual world to one in which Nick's chip is absent, and that is a distant world if for no other reason than that the chip's being implanted is a condition of the thought-experiment, or (b) just stipulate that a nearby world is one in which the chip is not present. I submit that when we vary the alternate/counterfactual conditions in Intervener cases, we can just

stipulate (b), so long as we keep track of the fact that we are varying the *ex hypothesi* conditions of the case. Option (a) threatens to warp the case in ways which will make it difficult to respect the relevant intuitions. Outside the phi-fi realm of Intervener cases, (a) is appropriate – in normal cases, distance to worlds is a substantive feature.

So, if we agree on (b), we subtract the chip, go to the nearest world in which Leno wants to do otherwise, and see whether Leno does otherwise. For it may turn out that removing the chip and entertaining a world in which the impulse comes over Leno to be honest reveals that Leno feels an irresistible greed at the prospect of an easy profit, and cheats Mrs. Pratt anyway. Suppose, for instance, Leno has recently tried crack cocaine, unaware of its instantaneously-addictive properties. Unbeknownst to him, in the actual world, the desire to cheat Mrs. Pratt seemed to have arisen spontaneously within him, but that sort of appearance of spontaneity often accompanies addiction-driven behaviors in people otherwise ignorant of the hidden motivational powers of addictive substances.

Thus, even if we remove external-agent Nick and his chip and restore Leno's alternatives, Leno may fail PAPW2. In this case, Leno lacks alternatives, but unlike Matthew, his lacking alternatives cannot be traced back to actions that satisfied PAPW2 that render him autonomous derivatively. For Leno's original consumption of crack was in ignorance of its instantly-addictive properties, so we *cannot* trace back his autonomy derivatively. If Nick knew what he was getting into when he first consumed the irresistibly-addictive drug, then his autonomy would be derivatively-traceable, together with his responsibility. Vicious addiction with foreknowledge is equivalent to virtuous self-cultivation with foreknowledge, but differs from indoctrination without foreknowledge – Stump's lapsed Baptist, who cannot, but wants to, have a drink (1993).

Baptist is not an apt target of praise, not even in her own post-reflective eyes, for her

virtuous, Luther-like aversion to alcohol is one she would rather be without, and fails PAPW2. Suppose she was like Matthew in having an invincible aversion, but one which didn't play a role in her actual-sequence decision not to have a drink, because she wasn't aware of that irresistible aversion. Suppose she just formed the desire, weighed it briefly against its opposite, and casually opted to refrain. This fails PAPW2, but satisfies PAPW1, just as crack-addicted Leno does when he appears to spontaneously cheat Mrs. Pratt, though such spontaneous impulses are often caused "bottom-up" from the still-hidden addiction. Baptist's PAPW1-satisfaction with her actual-sequence behaviors warrants the stamp of approval "*guidance-control*" on behavior she would rather be without, were she to try to have a drink. Baptist *would* feel unfree, though she may never discover this. Yet F&R would have to say she has guidance-control, F&R and Wolf would have to say she is praiseworthy, and Wolf would have to say she is free – since, for Wolf, what renders one responsible renders one free. Baptist is a strong counterexample to F&R and Wolf.

Consider a self-denying "Homosexual" whose gender-preference rationalizations, self-partitioning and other coping mechanisms have become deeply unconscious, inaccessible, and thus phenomenologically real – albeit false – indicators that he is not gay. Suppose he likes homosexual porn, has no attraction to women, and engages privately in homosexual behaviors, but is so homophobic he 'believes' he is heterosexual. His dating women is socially encouraged, and he participates – *apparently* willingly – in which case his behavior exhibits guidance-control.

If their analysis were correct, we would not hesitate to say Homosexual exhibits guidance-control. But if we can still coherently ask the following question, however, then F&R's equation of guidance-control with the fully voluntary will fail an open-question-type test: *But is his behavior really voluntary?* That question is coherent, given our inclination to

view his behavior as not fully voluntary. Thus, Homosexual's guidance-control fails the open-question test; Baptist and the crack-addicted Mechanic would also fail the test. Guidance cannot define responsibility.

Baptist, crack-addicted Mechanic, and Homosexual justify a demand for more than guidance-control or PAPW1-satisfaction, for weak regulative-control tested for in PAPW2 to rule out *the mere appearance of voluntariness*. Frankfurt would also insist on ruling out the *mere appearance* of voluntariness, but this sort of case explains the intuitive insistence on the further conditional ability, weak or determinism-friendly PAPW2. For in this case, *some* kind of further conditional ability above actual-sequence, PAPW1-satisfying guidance-control is needed to ground attributions of causal and hence moral responsibility.

F&R improve on Wolf, but do not establish the symmetry thesis; PAPW1 and PAPW2 do, and account for the warped intuitions in Baptist. Baptist poses a problem for Frankfurt, since she has a 2nd-order desire – alcohol-aversion – that compulsively prevents her from having the drink she, on later reflection, 3rd-order desires. Frankfurt's reply is that Baptist's will is divided; he insists on *wholeheartedness*, as Stump acknowledges (1993). But this seems unprincipled and weak; besides, stronger versions of Baptist may be adduced that satisfy wholeheartedness but still involve higher-order compulsions (Mele, 1995). If Frankfurt accepted the relevance of causal control, he could strengthen his reply with PAPW, which succeeds just where F&R and Wolf fail.

Something these cases share is the distance between possible worlds needed to apply PAPW. If Matthew's psychology is akin to Baptist's, then to locate the nearest world in which he is able to let the child drown or Baptist is able to take that drink, we may need to back-track to pre-agential points in their lives. If so, Matthew's bent is *not* traceable to a time in which he satisfied PAPW2, his saving the child is *not* derivatively free, and thus not

as praiseworthy, even if, in being a good deed, it is praiseworthy *as such*. In this case, Hero and Baptist are analogous.

F&R's guidance-control is insufficient for their purposes, but PAPW1 supports these more fine-grained distinctions. These cases illustrate that mere guidance-control, established by PAPW1-satisfaction, the truth-condition for F&R's jazzy guidance-control, is insufficient both for proving the symmetry thesis and for responsibility. For Mechanic and Hero are asymmetrical, and Baptist is not responsible for her unwelcome alcohol aversion, though F&R's guidance-control model renders them all responsible. Nor is Homosexual fully creditable for his restraint.

What establishes these facts is the distance needed to locate a world in which these agents satisfy the antecedents of PAPW2. If that must be located by branching from a pre-agential fork in these agents' life-paths, it defeats the derivative tracing formula for PAPW2. If so, they are not responsible, though they exhibit what F&R must concede is guidance-control. *This proves guidance-control is insufficient for responsibility*, though in *some* cases it appears sufficient, due to PAPW1-satisfaction. PAPW1-satisfying guidance-control is sufficient for weak responsibility; moderate responsibility is established by PAPW2. Stronger responsibility may established by PAPM, but the degree is left vague, as is PAPM.

There is a difference between guidance-control and voluntariness. Guidance-control is made out only intuitively through casuistry. Fischer, frustrated with his inability to substantiate the guidance-control intuitions in subtraction argument 1, appeals to the jazz analogy. But PAPW1 accounts for those intuitions in a principled way. There may be cases of guidance-control which fail PAPW1 and vice versa. Just because PAPW1 captures the intuitions in these cases doesn't entail that it is what F&R have in mind. Indeed, the cases they pick out may even be *coextensive* with the ones PAPW1 picks out, and yet the

intensions may differ. Still, whatever is picked out by their intuitive approach, PAPW1 will likely pick out, for a principled account takes up the slack of an unprincipled one.

In any case, PAPW1 isolates factors their vague account does not. The fuzzy/jazzy guidance-control they trade in cannot finesse the crack-addicted Mechanic, Baptist, and Homosexual cases, whereas PAPW1, being more fine-grained, can. But I am undecided about the directions open for interpreting what is meant by such context-relative notions as "nearest possible world" and related ideas. Let's explore directions for interpreting these ideas, both to uncover the complexities of these notions and to settle upon more or less plausible interpretations.

The reason we hesitate to call Homosexual's guidance-controlling restraint *really* voluntary can be explained by my account, but not F&R's or Wolf's. What explains the failure of the open-question test here is that it is not clear whether Homosexual satisfies PAPW1. His restraint *seems* voluntary, under his guidance-control, vaguely conceived, say, where some of the women he dates are sexually aggressive, but he declines, preferring to remain a gentleman. But the worlds we need to test PAPW1 are obscured by the facade of his heterosexuality.

To determine whether in dating these women, he would still do so *were he able to do otherwise*, we need to go far back into his homophobic upbringing to find the branch from which we may fork to a counterfactual Homosexual who is able to date men. And that may be a pre-agential point. For while the nearest worlds to which we will refer to test PAPW1 – on a superficial look – *appear* to be worlds in which he could do otherwise, since in those worlds he could refuse to date, come up with rationalizations for not dating that don't tip his hand, etc., the fact that he is able to not date the women in those worlds under those conditions is only partially telling about the voluntariness of his dating people of

same/different genders. The issue of intensionality in his intentions arises here, for how we frame his intentions depends on capturing the proper intensions; e.g., the categories of "women" and "gender he is attracted to", to name but two formulations, obviously cannot all be thrown into the same hopper.

It is folk wisdom that many heterosexual women feel comfortable around homosexual men, for there is no issue of sexual access. This sort of psychological fact needs to be taken into account with Homosexual, for just as gay men are not creditable for restraint in refraining from sexual aggression with women, neither is Homosexual. Thus, to capture the voluntariness or not of Homosexual's restraint, we need to go beyond worlds in which he has gender-preference-extrinsic reasons for being able to do otherwise than restrain, e.g., rationalizations, to worlds in which he has gender-intrinsic reasons for restraint, e.g., inability to be sexually aggressive with women. There, Homosexual knows he is gay and that he cannot sexually engage with women.

This location, by the way, does not tell the whole story about Homosexual. For here he is not yet able to do otherwise with respect to his 'gentlemanly' behavior, since, being gay, he is now *consciously* aversive to heterosexual contact. In this world, however, he is capable of dating men, in which case "able to do otherwise" is more thoroughly satisfied. That being so, moreover, sheds light on the intensional nature of his 'restraint', and debunks the previous interpretation of *self-control in the presence of sexual attraction* as *absence of sexual attraction*; only the former is creditable. Indeed, only the former exhibits restraint ability, for once Homosexual comes to grips with his homosexuality he *may* have no restraint abilities for men he is attracted to.

We must go farther out to a world in which he *can have heterosexual contact* to test for restraint, but in such a distant world he may not be recognizable, if gender-preference is

essential to personal identity. But suppose, before we go that far, we locate a world in which he processes his homosexuality through the lens of bisexuality, say, tells himself he is experimenting, in which worlds he is capable of female sexual contact. It may be that there is such an individual, if not many of them, who are primarily homosexual, but with some bisexual leanings, born primarily of their process of coming to grips with their homosexuality. If Homosexual is such, then we can locate a world in which he is capable of making sexual advances to women. This would be the closest world(s) needed to test whether Homosexual is genuinely engaged in voluntary behavior in refraining from making sexual advances to women, to fully test whether he satisfies PAPW1.

Were Homosexual a real person, it would be a contingent matter about whether he would or wouldn't behave differently once counterfactually able to do so, but we can stipulate either or each of the alternatives just to see where the intuitions go. So, we can describe his attitude toward heterosexual activity in this range of worlds as either positive, negative, or neutral, and possibly as varying from time to time. If heterosexual activity was only something he was inclined toward during the process of coming to grips with his homosexuality, after which time he lacked a bisexual or heterosexual interest, but felt no great aversion to women, then he would still be largely homosexual, capable of, but not inclined toward, heterosexual relations. Here, he would not be so creditable, though his refraining would be minimally voluntary, for in that world he would satisfy PAPW1 – he can make advances toward women, but refrains. But since he lacks interest in women, PAPW1-satisfaction doesn't render his behavior creditable.

In cases where he is bisexually active, though, say, thoroughly homosexual later on, he might make advances toward women given the internal/motivational opportunity. Here, say, he has interest in women, and either behaves as a gentleman or not. If he still refrains

from making advances toward women, then he is creditable for restraint, and his actual-sequence behavior is fully voluntary; if not, then his 'restraint' was a function of constraining factors having to do with his gender-preference restrictions, and was not fully voluntary.

This world may be too far from the actual world to support these intuitions, on the auxiliary supposition that gender-preference is essential to personal identity, and altering Homosexual's gender-preference excessively warps his identity. The supposition *that gender-preference is essential to personal identity* is an *auxiliary* supposition: Its truth is independent of the PAPW methodology, even if its truth/falsity plays a role in adjudicating these cases. For Baptist and other analogous cases in which it is necessary to backtrack to a pre-agential fork to locate the nearest world in which Agent is able to or wants to do otherwise for purposes of testing PAPW or PAPW also involve such massive warping of the person's character that I have taken it for granted that the intuitions fail in such cases.

If the world where Homosexual is bisexual is too far away for the test to be coherent, then the test may be run in the nearest analogous world, the one in which Homosexual knows he is gay, and is dating men. In that world, is he chivalrous? What we need to access to judge Homosexual's actual-sequence behavior is a world in which *Homosexual is sexually attracted to the person he is dating*, to see whether in that world he is creditable.

A crucial moral from the Homosexual case is that we need to be sensitive to features unique to each case to determine how to locate the relevant worlds for purposes of testing PAPW. This possible-worlds-determining function is highly context-relative, and blurred by the intensionality of intention/action. The differences between cases in which Homosexual doesn't make sexual advances to persons he is/isn't attracted to are partly intensional – dependent on how such persons are described relative to his interests. Analogous to the

distinction between event-*particular* and event-*universal*,[116] which depends on similarly-intensional differences between descriptions of the same general activity, disregard of which accounts for confusion in determining responsibility in Frankfurt-cases, disregard of the context-relativity and intensionality of the intentional components of these cases also accounts for confusion in determining degrees of voluntariness, autonomy, and responsibility. Issues of personal identity are also relevant, though also independent *auxiliary* matters.

Given the complexity of these matters, it is expected that simplifying assumptions threaten to distort matters. There needs to be a certain looseness in the formulation of interpretations for testing procedures for PAPW. These are features of the subject matter, not of the theory that purports to account for them. Insufficient attention to auxiliaries results in examples in which intuitions are contaminated.

PAPM functions as our all-inclusive criterion for voluntariness and weak conditional ability. Its application always engages auxiliary hypotheses extrinsic to the theory, which present theorists with equal difficulties *tout court*. That our treatment of these passed without defeating PAPM constitutes post-reflective equilibrium with our original intuitions. Our aim was to show that determinism-friendly-regulative-control, "easy autonomy", is established by PAPM-satisfaction, and that the metacausal analysis does a better job than competitors in handling intuitions hovering around Frankfurt-cases. Further refinements will be sketched in chapter 6, but our analysis has met its major aim. It takes the "semi-" out of the most-promising competitor theory, Fischer's semi-compatibilism. We do otherwise, on a

---

[116] This distinction may be made out with an example: In shooting the mayor, I may be responsible for the event-particular in which *I shoot the mayor at a certain time with a certain bullet*, etc., but not responsible for the event-universal *that the mayor is killed by gunshot wound*, if, for example, another shooter's bullet hit him in a vital area, killing him, whereas my wound may have killed him anyway, were he not also shot more fatally by the other shooter. Van Inwagen makes the distinction (1978, 1983); it is discussed in Fischer (2002).

daily basis, naturally.

With pessimism undermined at all three levels, we may now state with confidence that it is more plausible to think we have determinism-friendly autonomy, in all three ranges of cases, than it is to think that we do not. Having made our basic case, we turn to chapter 6.

Chapter Six: Applications and Suggestions

We can apply he metacausal theory to an error theory for inflated intuitions, e.g., contracausality. Our theory explains them as functions of phenomenological features of off-line mechanisms that *seem acausal*, but are *metacausal*. Subsumption under the causal aspect of the theory brings the phenomenon into contact with the natural order, deflates it, and reveals a recipe for how to correct anyone whose intuitions on the phenomenon led him to a nonnatural account of it. A natural account trumps a nonnatural one, *ceteris paribus*. Let us take a brief look at some of these tamable stallions.

Practical reason is a complex, off-line process that involves only weak possibilities. The off-line selection from, and on-line implementation of, considered RFAs is causally effective in forming and/or authoring one's choices/actions in ways that are immune from horizon-worries if one satisfies PAPM. Thus, practical reasoning is natural, contrary to those who place its operation outside the phenomenal realm, e.g., Plato, Kant, and the libertarians.

Contracausal intuitions require indeterministic underpinnings, and are held implausible for their metaphysical baggage. But the genesis of these notions is tied to the phenomenology of the off-line deliberator. In being off-line, he acts independently of 1st-order, on-line causal forces or motions. When off-line, he is *virtually* unmoved by 1st-order causation, and when he goes back on-line, he is *virtually* a *prime mover* of his acts. This is not all *virtual*: He possesses one *literal* form of contracausal power over 1st-order hydraulic causal dynamics, though there are both on- and off-line causes that ultimately move him.

This off-/on-line switching power differs enough from on-line/on-line 1st-order causal relations to warrant deflated talk of agent causation, without implausible metaphysics. That behavior is *ultimately* determined is irrelevant, *and it feels so*. The causally *disengaged* functioning of deliberative reasoning – though *engaged* in a *mediated* fashion – is the basis

for our error theory of intuitions of being able to do otherwise *under identical conditions*. The term "off-line" alone is suggestive of acausality, and invites a metacausal error theory.

It seems likely that we can take most of our cognitive/conative systems off-line, so we have what we may call "off-line muscles", a general ability like the ability to concentrate. Our use of it is likely the basis of inflated intuitions about freedom, as in the popular example of raising one's hand. Libertarians ignore that off-line causation is still causation. Their intuitions simply need to be bifurcated: Autonomy is housed in the off-line architecture, but the extra-phenomenal intuitions are byproducts of a dim view of its opaque features. Thus, a deflated view takes account of the fact that the 2nd-order phenomenology and causal disengagement of deliberation does not entail *total* causal isolation.

Social scientists, e.g., Meade and Perls, view the self as a socially constructed function that projects the judgments of others about the subject onto himself. Since projection is a form of simulation, so is self-formation, which may involve, e.g., simulating others' reactions to one's choices, reactions, and other possible behaviors. One need not adopt a social-construct theory to see the role of off-line processes in self-development.

The ability to simulate the characters one might become if one engaged in behaviors, and to lend one's will to some of these but not others, is a self-forming feature of autonomy. Ainslie speaks of lending one's will; Frankfurt speaks of identification of one's will with one's desires. In this process, the feedback-loop of teleological orientation, makes Ainslie's intra-agential bargaining possible, and enables a naturalized version of Kane's self-forming actions (SFA's) (Dennett, 2003) to be genuinely self-forming. In light of its on-/off-line and identification/alienation components, this account is immune to Wolf's (1990) self-creation *ex nihilo* objection. On "real self" or "deep self" views, Agent is free *iff* she acts from her real self and has a say in its development. Kane's is thus a real self view.

But real self views face a regress problem: To have a say in its formation, a self would have to exist before it is formed, but to have a say in that self, a prior self must first exist, and so on. Our theory is immune to this problem. Its models of distancing, metaconative identification/alienation and of naturalized SFAs render the self deeply causally responsible for its gradual creation. We may employ both intra- and inter-personal – off-line, metacognitive – simulations in every early SFA of deliberation; if so, we form ourselves in every significant early choice we make. If our earliest SFAs satisfy PAPM, then *we freely create ourselves* in our earliest significant acts of choosing.

Autonomy is required on some views of responsibility, e.g., the real self view and Kant's, which inflate the self. The belief in a self outside the phenomenal realm injecting novel causes into the causal stream (Chisolm, 1982) may be explained as a byproduct of an inflated view of opaque features of the phenomenology of off-line deliberation. But off-line disengagement, however opaque, does not entail *total* causal isolation. Our recipe is to show that the process only *seems* acausal because it involves off-line causation – no threat to freedom or determinism, thus no ground for burdensome metaphysical doctrines.

This analysis can be used in support of Plato's model of the ideal agent, who exhibits top-down control of reason over will, and will over passion. For Hume, conversely, reason is slave of the passions, so Frankfurt's unconflicted addict – a wanton, non-person – may only be criticized for lack of resourcefulness in maximizing hedonic returns on investments in heroin; Ainslie can criticize him only for poor intra-personal bargaining skills. For Plato, the addict is so malformed in his intentional structures that he functions as an animal, is unhealthy, and is spiraling toward 'self' destruction; for us, he lacks psychic homeostasis, i.e., autonomy. To Plato, the Humean model is only true of animals, the deranged, and other forms of wanton – not persons.

Watson (1975) adds a Platonic, value-based criterion to Frankfurt's, so he may make the same criticisms of Hume's view of the role of reason in the structure of the person. On Hume's view, self is an illusory cluster of fleeting sensory phenomena. It is no accident, then, that his model cannot account for the integration necessary for personhood. Hume's vision of the free agent is that of the unconstrained functioning of an amorphous aggregate of animal attractions and aversions, subserved by reason and cognition, unfolding according to natural law. But since there is no integrated self, there can be no 'agent'.

Nothing in Hume's account can provide for the possibility of persons, or differentiate them from animals. Ainslie's model also lacks conditions sufficient for an integrated entity that could play the role of a person. Thus, unconstrained animals must be free for Hume, and, on Ainslie's model, smart rats must be free. But on Plato's model, the will represents a broad range of attitudes such as indignation, and includes the moral sentiments and Strawson's (1962) reactive attitudes. Kant implicitly accepts a similar locus for the moral in the rational will.

Our model accounts for the integration needed for personhood, and rationally grounds our normative attributive practices. We can go beyond Strawson's account of these as rooted in ungrounded presuppositions of social life. For Intervener-style-self-regulative-control grounds normative ascriptions in *deep* causal authorship and thus desert, and accounts for why Agent identifies with itself and its values as functions of its own SFAs and higher-order considered preferences. The theory also offers a prescription for the attainment of ideal Platonic agency and suggestions for moral education, as we will see.

Social science construes norms as abstractions from collective attitudes, forming a 'generalized other', and individual values as internalized norms, formed by introjection from the generalized other. Simulation may play a role in both. Our theory thus not only *grounds*

our normative attributive practices, but can account for our *acquisition* of personal and social value structures. It thus can absorb those Platonic/Kantian views defended by Watson, Stump, and Wolf according to which Agent is free *iff* she acts in accordance with her real/deep values. We have a natural ability to simulate others' mental states, independent of the just-so story account of it, and are inclined to identify with them when we do. This grounds our normative practices, moral psychology, and the possibility of altruism. It grounds moral psychology by linking our ability to simulate others' credal-normative states to visceral sympathy; as Camus says, "to understand is to forgive". It grounds altruism by allowing us to experience how others feels, as if we are them (Goldman, 1993).

Developmental evidence suggests that to learn about others, we simulate them. As parents know, children learn most object-vocabulary through imitation of ostensive definitions, e.g. mimicry of sounds associated with items. They *get it* that the noises and gestures their parents are making have an *aboutness* to them, that their parents are *trying to convey it* to them, and thus that there is a *content* to their parents' speech and gestures. It is likely that this process *already* involves simulation: The child tries to simulate the parent's mind in looking at x and uttering x-related noises – thanks to our primatological penchant for both gaze-following and its adjunct, mimicry.

Parents link exaggerated sensations in pitch, facial gesture, etc. with disapproving attitudes towards "bad" items, to guide the child. This primatological behavior, e.g., displaying of teeth, raised brows, etc., is the "mother's knee" basis of the child's acquisition of the reactive attitudes; recall primate emotivism. The developmental data support the claim that learning about normative expressions capitalizes on primatological gaze-following and other mimetic skills. Since learning others' intentions is a precondition for forming the full range of reactive attitudes (Adler, 1997), and simulation is a precondition for such

learning, simulation plays a key role in the acquisition of our normative attributive knowledge, both of the reactive attitudes of others and of ourselves.

In simulating others, we mock-experience what makes them like us, and see differences as functions of different intentional inputs. If we adopt their input stance, we get their perspective on things. This is why we develop common attitudes under similar circumstances, and it grounds the ethical axiom that *like cases ought to be treated alike.* Hume held that we are given to sympathy viscerally, and thus there is no rational basis for our emotive attitudes and judgements. Strawson claims that the three kinds of reactive attitudes, self-, other-, and vicarious-reactive, stand and fall together (1962), but are brute facts about us. They are both wrong: How we react given similar input is the same for each of us, so our ability to simulate others or ourselves under similar or different input is what accounts for sympathy and the intimate links between the three types of reactive attitude.

On our theory, our normative practices may be grounded *factually* and *normatively.* Their *factual* grounding is achieved by locating their genesis on a late segment of the evolutionary continuum reflected in my causal column. That segment begins with gaze-following and mimicry, and proceeds with primate emotivism. Primate emotivism describes social expressions of approval/disapproval as functions of natural selection pressures in primate social life. In us, these evolved into reactive attitudes and other sentiments, and may involve such sophisticated mechanisms as counterfactual imagination (Walter, 2002), described in the rude motorist case. This explain the factual basis of resentment. The *normative* grounding of the attitudes refers to the idea that simulation presupposes the belief that persons are importantly alike.

In many parts of our lives simulations are conspicuous and essential. As Strawson claims, life as we know it would not be the same without the network in which the reactive

attitudes are intelligible. Diagrams, models, and illustrations of any form are all simulations. While they are not *logically* necessary in math or theory construction, they may be *psychologically* necessary for understanding. Consider the prominence of imagination in Kant's cognitive architecture. In addition, toys and games involve simulations, as do acting, theater, and literature. Consider also that celebrity actors and musicians – high priests of modern culture – are athletes of simulation, which might be a reason they are held in such high regard. While *imagination* is essential to the arts and humanities, *supposition* is necessary for philosophy, mathematics, and science. Supposition involves a form of simulation (*cf.* Sorenson (1992) and Stalnaker (1984)).

In light of the possible role of simulation in language acquisition, the Gricean communication model may be supported if a speaker's simulation of the listener is involved in anticipating the effects of the speaker's words on the listener. Grice suggests we pick utterances on the basis of predictions about their effects on their hearers' mental states. Simulation fits nicely: The speaker forms expectations about the hearer's mental states by 'seeing' how he himself would react to hearing certain sentences, via a simulation along the following lines. Speaker S runs utterance U1 and perhaps hearer H's belief set B1 through S's off-line language processor and related mechanisms, which produces mental state M1 in S, which is not what S wants to occur in H, so S runs utterance U2 and B1 (or B2) through, etc. More generally, however, language itself is essentially the substitution of phonological or inscribed stand-ins for objects, and since stand-ins are simulations, language involves simulation *essentially*. If so, it would be second-nature for us to juggle Gricean simulations

in even the most complex speech and thought.[117] This simulational feature of communication would connect our normative attributive practices with our very use of speech and thought, a serendipitous link with social survival value.

Our account adds to Dennett's (1984, 2003). The *Sphex* wasp is tropistically determined, but we are not an opaque collection of tropisms. My account tries to make more explicit than Dennett does precisely how our flexibility emerged. There are causal-column watermarks for flexibility in the evolutionary record. Plant behavior is determined by, say, "soil causation", i.e., roots fixed in soil. Animal behavior is not; animals move. Imagine plants that uproot, move, and feed off various soils. Such locomotory plants would be less soil-determined, more "free". Like *highly* locomotory plants, we can block off 1st-order causal inputs and redirect their momentum within our metacausal system; animals cannot.

Determinism must be compatible with locomotion, or locomotion would disprove it. Since autonomy is an abstract locomotion ability, it must be compatible too, or else daily activity would disprove it. Just as locomotory life forms are less exogenously determined than their fixed counterparts, metacausal agents – Flex – are much more endogenous, for they can pick up their 1st-order causal roots and operate metacausally; *Sphex* cannot.

These differences suffice to account for the differences in autonomy between average members of the two species, and for the intuition that there is a graded autonomy scale for species. Certain animals, e.g., primates, have linguistic and related abilities, and their degree of autonomy seems closer to ours than to animals who lack such abilities, e.g., reptiles. We may classify species as free to the extent that they are metacausally differentiated.

---

[117] Since speech involves simulation, it is metacognitive. *Most* thought is linguistic; Ryleans view it as speech *sotto voce* (Dennett, 1984); it is also *representational*, involving a *stand-in* for the original *presentation*. In either case, it is simulational and metacognitive.

Reason seems post-evolutionary. This invites an inquiry into the status of reasons. We say that we do the things that we do *because* we have such and such reasons. Though the "because" here is partly normative in that we *rationalize* our actions by reference to our reasons, it is also partly causal, for our reasons are what lead us to act. Since agents act on reasons to bring about desired ends, however, this is a teleological form of causation, a type of causation most hold implausible. The problem is, if reasons are genuinely causal, then they must be nomological. And if reasons are causes the lawful effects of which are actions, then reasons are like desires – mechanical, robotic, involuntary: Reasons mechanically push people, as desires mechanically push animals, into action – an image that invokes CON.

Davidson's anomalous monism provides an escape (1968). There are no laws that relate events described as 'catastrophes' and 'hurricanes', though these admit of lawfully related descriptions at more basic levels. The level of description at which behavioral events and their psychological causes are described as actions and reasons admits of few or no laws that relate them *as such*, though all the elements of these phenomena admit of nomological descriptions at the neural and more basic science levels. But neural description cannot capture the normative or pragmatic elements of reasons *so described.* There are few or no laws that relate reasons and actions as such, but that does not entail that reasons-based behavior is acausal or nonphysical; all that follows is that it is anomalous *at the level of description* at which it is rational.

Just as the metacausal model supplies details for Dennett's functional complexity account, so, too, it helps fill in some of the details of anomalous monism: The nomic architecture of the mental is not merely causal, but *metacausal.* But the unique psychoneural pathways that instantiate agential metacausal structures vary from agent to agent, so I call its nomic character "idionomic". In short, just as *metacausal* doesn't entail *acausal* and

anomalous monism doesn't entail *anomie*, *idionomic* doesn't entail *anomic*.

Wolf's "Reason View" of freedom (1990) does not address issues in philosophy of mind. She implicitly presupposes a Davidsonian view of mind, and contrasts her view with "Real Self" and "Autonomy" views. One of her insights is that *incompatibilists of all stripes* presuppose that responsibility requires *autonomy*, whereas compatibilists do not; she cites Kant and Sartre as exceptional on extrinsic grounds.[118] Wolf presupposes that autonomy leads to incompatibilism, but since ours is a *compatibilist autonomy view*, it contradicts her claim and helps identify what is unique about the metacausal theory.

On the Real Self View, Agent is free *iff* she can act in accordance with her 'real self', i.e., the intentions that issue from her *valuational* – not just her *motivational* – system (Wolf, 1990, p. 75, citing Watson, 1975). The autonomy view considers both systems insufficient:

> To be responsible, they say, [Agent] not only must be able to govern her actions by her real self, she also must be able to ensure that her real self is not in turn governed by anything else. (Wolf, 1990, p. 76)

Wolf's main objection to the Real Self View is similar to the one made by the autonomy view: Agent's values may be constrained, such that her having the values she has may not be up to her at all. The Reason View substitutes an ability more general than that involved in Agent's ability to act according to her valuational system, what she calls Reason, the highest of all faculties, which consists of the ability "to 'track' the True and Good in her value judgments" (Wolf, 1990) and act accordingly. Wolf states,

> [T]he difference between the Real Self ... and the Reason View [is that on] the Real Self View, an individual is responsible if and only if she is able to form her actions on the basis of her values. The Reason View insists that responsibility requires something more[:] an individual is responsible if and only if she is able to form her

---

[118] Contrary to Wolf, however, Sartre is an incompatibilist, for he is widely taken to advance a thesis of radical freedom. Moreover, Kant's 'compatibilism' strains the meaning of the term, and does not mean determinism is compatible with freedom *in the physical world.* So, *all standard* forms of compatibilism reject autonomy as a condition of responsibility.

> actions on the basis of her values *and* she is able to form her values on the basis of the True and Good. (1990, p. 75)

There is enough of a difference between Agent's values not being "governed by anything else" and her being "able to form her values on the basis of the True and Good" to warrant the idea that the Reason View is distinct from the Autonomy view, but not enough to block the idea that the Reason View is just a modified version of the Real Self View. Saying that one has an ability to form one's values on the basis of Reason is just another way of saying that one has the ability to form one's values on the basis of Value, or one's reasons on the basis of Reason. The contrast between Reason, value, and autonomy needs to be examined.

But being *able to* form and act upon values on the basis of Reason is an autonomous ability, so Reason provides at least one way to *realize autonomy*, and may be subsumed under autonomy. Similarly, if merely acting on the basis of one's values is insufficient because one's values may be constrained, then merely forming them on the basis of Reason is insufficient as well, because tracking and acting in accordance with the True and Good may be constrained. There is no reason to hold the Reason View superior to the Real Self View either. There is *no* parallel sense, however, in which *autonomy* may be constraining.

Wolf mischaracterizes autonomy as ability to act for *or against* Reason, to make radical choices, to act irrationally or on no basis at all. She rejects autonomy, for if two agents possess the abilities required under the Reason View, but only one possesses autonomy, that doesn't qualify her for greater praise or blame. What Wolf fails to see is that what is preferable about autonomy is Agent's ability to determine what will be the basis of her own reasons, values and actions. Wolf is blinded by her persuasive definition of Reason:

> ["Reason" refers] to the highest faculty or set of faculties there are ... that are most likely to lead us to form true beliefs and good values. If [so], the autonomous agent must be one who is able to act in accordance with Reason *or not*. We have seen, moreover, that this ability to choose among the rational, irrational, and nonrational

> alternatives alike is not an ability to choose on some higher-than-rational basis. Rather, it is an ability ... to choose whether to use any basis for (subsequent) choice at all. (Wolf, 1990, pp. 54-55)

Wolf seems to rule out, definitionally, a faculty higher or more general than Reason. But a metacausal volitional system may be construed as superordinate to Reason, since *any or all of the processes that realize 'Reason' may be taken off-line by an act of will.*

Wolf dismisses Frankfurt's analysis because "an agent who is alienated from her first-order choice may be alienated from her higher-order choice as well" (1990, p. 30). But Agent may equally be alienated from her Reason. Consider a Manichean insistence on only performing maximally justified actions or accepting maximally evidenced beliefs. Such paradigms of Reason would be candidates for therapy. Wolf bars all counterexamples to the Reason View by speaking in the abstract about Reason as the *highest of all faculties* or as *whatever* tracks the True and Good, but this cannot vacuously mean *whatever is invulnerable to counterexamples.*

Imagine a case where an ideal True- and Good-tracker's Reason is manipulated by manipulating her environment (Mele, 1995). Wolf may consider the chains of Reason golden, but they may be chains nevertheless. Here Agent has an interest in acting against Reason – unless that sort of agent is *defined away.* Wolf tries just that, for her reply to these sorts of counterexamples is to weld Reason to "normative pluralism", which holds there is not only one set of normative truths. If so, it is contradictory to 'track' the Good or maintain that it is a subset of *the True*, as she says it is, or that Reason is the highest of such faculties: Either the tracking model collapses on pains of 'alethic pluralism' – the claim that there are incommensurable sets of truths – or else a reason to act against Reason must be glossed over by making Reason so pluralistically inclusive and attenuated that appeal to it borders on vacuity. Neither will do.

Even if we grant Wolf's inclusive view of Reason, not all behavior within the sphere of our normative attributive practices is a function of True- and Good-tracking. Love, for example (G. Strawson, 1986, Frankfurt, 1999), can involve elements of Reason, but often is motivated by factors that have nothing to do with Reason, however construed. Yet we regard it as an appropriate object of regret/praise. An artist's motives in creating a piece may be *arational*, gripped by a Muse, or carried away by an emotion or stream of consciousness. The same arational factors may motivate much behavior that is facially Reason-based; someone may be attracted to the virtues for visceral, unconscious purposes ultimately explained in evolutionary terms, by aesthetic sensibilities, or in other arational terms.

Indeed, it is a Muse-invoking strategy of the creative to circumvent Reason, and *having a reason to circumvent Reason* doesn't subsume Reason-circumventing behavior under Reason.[119] To say, *ex post facto*, that such *arational* indulgences typically pay off for artists, lovers, or the virtuous and so are justified and covered by pluralistic Reason, as ways of "tracking the Beautiful", the "Mystical" or the "Erotic", either begs the question or misses the point: Love, art, and other amoral and/or arational activities in the core of our normative attributive practices are *mis*characterized by the Reason View, though not by the Autonomy view. Wolf treats *arationality* as if it were in a league with *irrationality*. While *irrational* behavior has little to recommend it, we need autonomy for *arational* behavior.

Thus, to characterize autonomy as the ability to be *irrational* or *insane* is to erect a straw man. On a charitable interpretation, Reason is extraneous: Autonomy is the ability to do or refrain from doing any act, to let any intention lead to action or refrain from doing so. Likewise, ambidexterity is the ability to use both hands in equally skillful ways, not the

---

[119] See Dennett (1984), pp. 65-66, Parfit (1986), pp. 12-13, Mele (1987), ch. 10, and Pears (1984) on useful irrationality; see Mele (1995), pp. 177-194, on irrationality in the face of CNC.

ability to use them according to Reason or not. Autonomy is simply a *doing*-ability more inclusive than dexterity, the *most* inclusive, that includes the ability to place any number of cognitive/conative systems on- or off-line. Reason-based cognitive systems are one subset of such systems. There are reasons for wanting autonomy, but not every reason for autonomy is, *qua* reason, a reason *for* the Reason View: We want autonomy for itself and because it constitutes what it is to be a person. What one does with one's autonomy is definitive of individual character. Nothing about these reasons for wanting autonomy involve "the True and the Good". And unlike responsibility, which is at least partly nonnatural or normative, autonomy is a strictly natural, metaphysical matter.

Love and art involve the arational; so do the creative, emotional, and erotic. "Emotional intelligence" competes with Reason (Goleman, 1995), as do the mystical and transcendent. We may construct counterexamples to the Reason View from these domains of "the Arational", if, with Wolf, we capitalize. We may *reinterpret* these to *comport with* Reason, but their value is missed by the Reason View. Wolf's theory is too normative:

> Since the point of the Reason View is to stress that the kind of freedom we need for responsibility is a freedom within Reason, a freedom to be governed by Reason (or, if you like, Rationally to govern ourselves), it may be thought that, insofar as Reason fails to pick out a uniquely best choice, the maximally free agent must be able to choose among the Rationally acceptable choices independent of all forces whatsoever. *Such an understanding, however, seems to me to be a symptom of a continued tendency to see freedom and responsibility as ultimately metaphysical matters, and not, as the Reason View proposes, as more primarily normative ones.* (Wolf, 1990, p. 144; emphasis added)

Either Wolf's view is *eliminativist*, so freedom is unnecessary for responsibility and thus does not exist, or *entirely normative*, so that freedom just is *whatever* responsibility requires. Her freedom is determined by her ethics – a reverse is/ought error. But even if freedom is defined by responsibility, that turns out to require autonomy. Again, to avoid manipulation and manage the Arational, autonomy is necessary *and* desirable.

Wolf's normative model fails to distinguish between unconstrained/constrained Reason-based action, and this is primarily a causal/metaphysical matter. If one *must* act in accordance with Reason, one is *unable* to do otherwise. One is *fully* free *only* with respect to things one is able *not* to do, even if one is *responsible* for what one *must* do. Ability to act on Reason is *one way* to develop ability to *not* act on the basis of impulse, a form of *autonomy* – control over the bases of what one acts on.

Wolf distinguishes between superficial/deep responsibility: The former is only causal; the latter is moral. Wolf adds that many athletic, intellectual, and artistic endeavors admit of a similar distinction, though she focuses on the moral. But the distinction may be expanded to include non-morally deep attributions. A skilled artist and a child may each produce art that appears in all relevant respects alike, but only the former is *deeply* attributable to the author. Similarly, merely acting on one's values may be a superficial expression of self, whereas forming one's self on the basis of sensitive reflection may be an expression of deep-self behavior. Other areas admit of a similar distinction, such as friendship, parenting, and teaching. Why not include gardening, if not all activities? Anything we do may reflect on us superficially or deeply, depending on the circumstances.[120]

Strawson (1962), too, has gone beyond the moral in addressing freedom, and has identified the reactive attitudes as including our visceral and emotive reactions, and our normative attributive practices in general. He held that any rational reconstruction that accepts the consequences of (CON's) determinism – e.g., one that would exculpate all behavior – would either be practically impossible or render whatever is left of human life unrecognizable. For the normative beliefs which underpin the reactive attitudes are so

---

[120] In light of exclusionary remarks Wolf made (1995), she would likely deny this.

deeply entrenched in our way of life as to be indispensable to it. The metacausal theory bars global exculpation, and maps onto our normative attributive practices in ways that rationalize them consistently with determinism.

The reason we hold 1st-degree murderers maximally culpable is that they deploy such higher-order deliberative-mechanisms as prudential planning and egoistic strategizing, conditions sufficient for *mens rea*. The murder is not an immediate result of lower-order on-line causal antecedents, such as the emotions that might be present in aptly-so-called 'crimes of passion'. Since a free act is authored at higher levels of the metacausal structure, and murder is of just that sort, it is free. Our deep negative reactive attitudes are grounded in the fact that murders are *deeply* attributable to murderers as autonomous agents.

By contrast, stepping on someone's shoe while being shoved onto a crowded subway car usually results from a form of causation that involves minimal higher-order processing, and that is why it is partly excused. Stepping on someone's shoe on a crowded train is not always excused, depending on circumstances. If the pressure from behind is closer to that of a rampage, then avoidability is lessened: This is closer to the sort of motion-transfer causation true of mid-sized objects in Newtonian physics. It is thus no surprise that we do not develop resentful reactions to the offending parties: There is tacit recognition of the sort of non-negotiable causation that triggers an exculpatory response. Conversely, if there are only a few people eager to get into the subway car, and the pressure they exert on one from behind is slight, avoidability is higher; consequently, exculpatory responses are not evoked, but resentful ones are more naturally invited. Stepping on someone's shoe in cases of very low avoidability originates almost entirely at the 1st-order, on-line level of causation, whereas in cases of high avoidability it may originate from some higher-order off-line level.

Reactive attitudes and associated normative attributions, then, map onto the

metacausal architecture just where the theory predicts them. So, e.g., the kleptomaniac acts from lower-order-dominant causal forces (compulsion), with little or no self-control, and so her act elicits some sympathy, just as the theory predicts.

We may draw the domain of the reactive attitudes further within a rational context by reference to simulation. Thus, we may know others' reactive attitudes if we simulate their credal-normative, emotive, and visceral systems the way Goldman (1993) argues we ground interpersonal utility judgements in a simulation-deploying epistemology. Indeed, these are related: By simulating hedonic states and evaluative weighting mechanisms to arrive at interpersonal utility judgements, we are manipulating a credal-normative reasoning mechanism akin to the one responsible for the generation of the reactive attitudes.

Our account makes possible a rational reconstruction of our normative practices, and is sensitive to determinism. It does not violate Fixity, nor lead to the dreaded results of global exculpation, the ending of all attitudes and normative practices – Strawson's fears for any determinism-sensitive reconstruction. We have identified those conditions that do and do not underpin *local* exculpation, reactive attitudes, and normative attributions. No global or catastrophic consequences are so much as *suggested*, so there is no peril to our way of life. Strawson suggests our normative practices may be severed from metaphysical questions about the causal-nomological nature of action, but these issues are happily related here. Thus, Strawson is prematurely pessimistic about the threat from determinist metaphysics to norms. Since freedom is unproblematic, his pragmatic reconciliationism is unwarranted.

Our theory – unlike Wolf's – is immune to a determinism-sensitive Strawsonian objection about asymmetry in praise and blame. The Reason View holds a person free and responsible even if *unable* to *fail* to act on Reason, so long as she is somehow 'able' to act on Reason. Let's call "actualistic ability" an 'ability' to do x but not to *not* do x. It is the ability

that is present when PAPW1 is satisfied but PAPW2 is not, and appears in Locke's prisoner and Frankfurt-cases where Agent's 'free' choice lacks alternatives.

But it *only* makes sense against a background assumption of full, PAPW2-type agency that includes the ability to *avoid* the fortunate choices. We only hold actualistic acts responsible because it's *stipulated* that Agents would have so chosen *were alternatives present.* Thus, even if we hold that they are responsible and they could not do otherwise, this does not warrant divorcing responsibility from full autonomy. For it is the tacit PAPW1 idea that they would *not* have chosen otherwise *were they able to* that justifies ascriptions of voluntariness and thus of responsibility. For it is stipulated that *the motivational state is identical* in agents who would not have chosen otherwise *but could have*, and agents who would not have chosen otherwise *but could not have.*

Hard determinists, libertarians, and semi-compatibilists think determinism entails all we could ever have are actualistic abilities. But our arguments suggest otherwise. The fact that *some* – exceptional, phi-fi – behavior is actualistic doesn't generalize to even *much* behavior. Rather, *most* human behavior involves PAPW2-satisfying, non-actualistic conditional ability. This "Privation Subtraction Argument" subtracts the limitations from privation or *actualistic* cases as ones that fail PAPW2 and infers the properties of fully functional or *conditional* cases as cases of PAPW2-satisfaction.

Again, agents on the Reason View are *praiseworthy* for doing what comports with the Good but *not blameworthy* for failing to, *though both are actualistically determined.* Wolf tried to evade her own asymmetry on the grounds that agents who *should have* known better are responsible, but this is unconvincing, and *ad hoc.* For whether they know better or not, if

they cannot act on Reason, they should *not* be held responsible, despite what Wolf says.[121] If agents know and *can* act on Reason, but do not, then they are *not* actualistically determined, they possess weak conditional ability, and their autonomy grounds their blameworthiness. She could say that on an akratic model, stronger motives overpower Reason, so we should forgive the actualistically helpless agent, but not all blameworthy behavior is akratic. But agents aren't really helpless, since they are not *only* actualistic. Our theory is immune to asymmetry: Agents are responsible to the degree that they can control what they do.

Let us return to the 'error' part of the error theory. The phenomenology of our ability to *toggle the psychoneural gates* that dis/engage our cognitive/conative systems, "gate switching", I argue, inflates our intuitions about freedom. When we take intentions off-line that would issue in actions on-line, and put off-line intentions on-line *at will*, it feels *up to* us, as if we are not moved by any force, though we move the world. We imagine possible worlds tied to intentions, select one, and go on-line with it; we can simulate intentions we don't have, cultivate ones we want to have, and eliminate others. In doing so during self-formative stages, we feel these abilities enabling us both to design ourselves and author our futures. We seem to be self-created unmoved movers.

As self-designing authors, we *seem* uncolored by nature. Our freedom suggests the power to do otherwise under *identical* circumstances. The error is to infer from *the phenomenology of weak regulative-control* that the exact elements of choice could be held fixed while another could have occurred, for this ignores the causal power of the conative details: To think an agent who went on-line at T could have stayed off-line at T *under*

---

[121] Acting on Reason or not is analogous to a weathervane responding to the wind or not: It is not up to the weathervane whether it is rusty or not; by analogy, it is not up to Agent whether he acts on Reason or not, on the Reason View. Thus, there is no basis on the Reason View for any asymmetry. Sober's weathervane theory (1995) is also insensitive to this problem.

identical circumstances ignores that her going on-line is a function of her wanting to *just then*. If agents want to go on-/off-line, but cannot, that is akrasia – not just determinism. *Akrasia* is a *privation of autonomy*, not *the default* for determinism; only *autonomous* beings *can* suffer *akrasia*, and autonomy is not opposed to determinism: The fact that our autonomous agency is *ultimately* determined by prenatal elements is irrelevant, for there is enough autogeny/self-authorship to 'cancel out' the pre-metacausal antecedents.

The error is to assume that the mechanism that subserves the phenomenology of choice is contranatural, but metacausality is a species of causality. Yet it is functionally-complex enough to explain why it is taken as contranatural: Its relatively-endogenous self- and action-controlling powers are 1st-personally transparent, but its exogenous elements are phenomenologically opaque. It is tempting to go contranatural here, because we fear our normative institutions will fall unless we are ultimate causes. The idea that *it is ultimacy or nothing* is what sustains incompatibilist optimism. And indeed the ultimate causes of behavior extend back to before our births. Autonomy guarantees origination (O'Connor, 1995), but is thought incompatible with determinism.

Metacausation is monistically causal, *ergo* compatible with determinism, but anomalous at the level where the principles that govern practical reasoning involve normative rules. There are no *purely causal* laws at this level (Davidson, 1968). This anomalism is physicalism-friendly, for though phenomena folk-psychologically described are not type-type reducible to those described neurally, principles of neuropsychology can account for how the brain instantiates rational structures (Walter, 2002, pp. 569-70). PAPM-satisfying non-epiphenomenal mental causation is consistent with token-token physicalism, since the active causal elements function at the *highest*-order physical or *meta*causal level (mind), as opposed to the *lowest*-order physical or causal one (brain). This naturalizes

autonomy: The ghost *is* the machine – a self-soft-wired, highest-level psychoneural *functional integration* constructed by the "idionomic" metacausal structure. The theory specifies the causal/functional features of act-origination, agent causation, and the attribution of causal authorship, in a way that rehabilitates inflated libertarian intuitions.

There is an element of autonomy *qua Kantian self-legislation* in the adoption of an idiolectical credal-normative system and in its associated self-sculpting process. The self-system idionomically governs how its intentions will count in its deliberations and thus which RFAs become what Davidson (1968) calls "primary reasons", i.e., those RFAs that lead straightaway to action. The structure of one's "idio-system" is soft-wired in one's psychoneurally-realized metacausal structures, individuated in the SFAs (Dennett, 2003) of early development. Each such element is subserved by a unique neurosignature in the higher-order functional networks of the mind-brain, and parallels that of idiolectical memory associations from unique experience. The control-levers for these 'gated' psychoneural paths are in reflexive consciousness. These structures are complex and involve elements that appear anomalous *when viewed across levels*, but that does not imply that metacausality is *anomic*: These idionomic psychoneural relations are subserved by lower-level neurophysiological states that admit of nomological descriptions. 1st-order causality is nomic, and metacausality is *idionomic* – not *anomic*. Each of us is responsible for the SFA-generated sculpting of the "idio-structures" that subserve our practical reasoning system, so our actions are significantly endogenous, under our control thereby.

The libertarian wanted acausality because he thought we must originate our actions to be responsible for them. But Agent is a plausible originator of actions, for their determinants are significantly shaped by his own self-forming and -ruling metacausal psychology. Hence, he is metaphysically responsible for them, and there is no need to try to escape determinism.

Pre-natal ultimacy is impossible, but it is a red herring.

Similarly, the theory fills in details of Searle's model of mind (1992) in a way that reveals how incompatibilists are led astray. His view is that mind is an ineliminably subjective higher-order neural state; this amounts to noneliminativist materialist monism – despite Searle's own antipathy for such traditional categories. Yet, despite his defense of the mental as higher-order physical, he thinks lower-order causation (or micro-physical explanation) delegitimizes freedom.

> As long as we accept the bottom-up conception of physical explanation ... on which the past three hundred years of science are based, then psychological facts about ourselves, like any other higher-level facts, are entirely causally explicable in terms of and entirely realised in systems of elements at the fundamental micro-physical level. Our conception of physical reality simply does not allow for radical freedom. (1984, p. 98)

Searle implicitly shares Wolf's contranatural misconceptions about autonomy as a radical ability to act without cause, motivation, or Reason at all; we have shown this view to be inflated. But just as the view of mind as nonphysical and thus incompatible with body has led physicalists to eliminativism, so too the view of autonomy as contranatural and thus incompatible with nature has led physicalists to reject autonomy. So, just as eliminativism is unwarranted in light of higher-order physicalism, so, too, the rejection of autonomy is unwarranted in light of higher-order physicalism. Again, if mind *just is* physical, then autonomy *just is* physical. It is plainly asymmetrical and illogical to rob the mental of causal efficacy if the mental *just is* physical.

If Searle's bottom-up formula *does* undermine the legitimacy of higher-order explanations or causation generally, then the same formula would eliminate the legitimacy of the "irreducibly subjective ontology of the mental" as well, a major theme in his recent work on mind (1992). Further, the idea that freedom *just is* a higher-order physical property seems

to be an almost-immediate consequence of Searle's analysis of mind. If there is nothing in the bottom-up causal model that contradicts the legitimacy of the mental as a nonepiphenomenal higher-order physical state with causal properties, as Searle conceives it, then freedom can be construed simply as the subset of those higher-order causal abilities *the control of which is located at the higher-order level.* We need not sacrifice the ontological legitimacy of freedom to bottom-up reductionism.

Bottom-up physicalism is a dogma that may be vacuous. Consider the *philosophical* claim – it is not a claim based on experimental evidence – that quantum indeterminacy 'cancels out' at the macro-level, so macro-level determinism is determinism enough for purposes of incompatibilism. This inconsistent claim is a sheer assertion supported only by the eminence of bottom-up physicalism. But *macro-determinism* contradicts *bottom-up* physicalism – it cannot be *both* canonical in, *and* contradictory to, bottom-up physicalism.

More importantly, what is good for bottom-up physicalism is good for anyone else. Thus, if the notions of *uncaused micro-event* and of *micro-event the causal effects of which cancel out at the macro-level* are licit, then we can infer *that there are uncaused events* and *that micro-level causation can cancel out at the macro-level.* From these we can infer *that bottom-up neural causation can 'cancel out' at the macro-level of mind*; that metacausation can be functionally cut off from lower-order causation by a canceling process, thus indeterminate *relative to* the lower-order level; and that metacausation can be indeterminate *relative to* micro-causation, as when Agent literally engages contracausation relative to lower-order causal forces he disengages. Nothing about energy transfers across levels that involve a switch between merely pseudo-indeterministic/deterministic modes (either way) entails contranaturalism.

A general metacausal principle is that the higher the metacausal level from which

actions issue, the lower the degree of animal causation, constraint, or unavoidability, and the higher the degree of autonomy, personhood and responsibility thereby. This "corollary freedom hypothesis" has a broad range of empirical and practical consequences. For instance, it predicts certain relationships to hold, say, between highly metacausal states and liberating experiences. Thus, consider the tremendous top-down metacausal restructuring (reason and will over passion) needed to transform oneself into a skilled athlete.

Joggers experience "flow" and "runner's high".[122] Their neuro-muscular reflexes are developed under the ceaseless exertion of their higher-order wills over lower-order tendencies to yield to pain and fatigue, with environmental and biological feedback systems oriented toward the self-system's autotelic ends, with one-pointedness of mind: It is no wonder they enjoy liberating psychological states.

Similarly, certain authors, scientists, mathematicians, chess masters, and others may be considered to have highly differentiated metacausal structures, since they possess skills at bringing to bear a large number of symbolically-ordered principles and rules on a tremendous mass of information; so, too, for adepts at ostensibly non-intellectual tasks, such as artists and mountain climbers, so long as they are in active command of a vast body of skills and related knowledge. The mastery of these complex bodies of knowledge and related skills yield not only liberating experiences, but higher levels of psychic integration; since these autotelic individuals are paradigms of metacausality, the predictions of the corollary freedom hypothesis are confirmed in the link between their peak experiences and their metacognitive sophistication.

Most dramatically, yogis, zen masters, and other paragons of top-down metacognitive

---

[122] Studies suggest they can toggle the psychoneural gates for pain and endorphins, the brain's pain relievers, which combination engenders euphoria (Csikszentmihalyi, 1991).

training vividly illustrate the empirical consequences of the corollary freedom hypothesis, for these individuals report a great variety and intensity of ecstatic, transcendent, and otherwise liberating experiences (Austin, 1998; Benson, 1975, 1996; Goleman, 1988). What is striking cannot be explained away is that those who have them are *radically transformed* thereby. Even Nozick seems to admits this at the conclusion of his soundless stereo argument:

> Will this debunking explanation have more difficulty in explaining the surprising and often momentous changes in the people who have the experiences? (1981, p. 159)

They discover renewed meaning, purpose, and spirit; they rearrange their lives; and they become altruistic to the extreme. Such radical transformations of perspective, character, and purpose constitute solid abductive evidence for the corollary freedom hypothesis.

These transformations may involve a shift above a previous threshold to a higher-order situs of functional integration that yields greater metacausal control. *Zen* masters and other metacognitive adepts speak of a radical figure/ground shift in perspective, from *identification with* the stream of experience, to *detached awareness of* the stream of experience. As Goleman puts it, "[i]n this state, [meditative awareness] crystallizes as a constant mental function" (1988, p. 96). A figure/ground apperceptive/conative shift *from cognitive to metacognitive identification* would free one from the lower-order contents of one's mind; if so, stimuli, intentions, have less power to pull/push one to/from behaviors. Psychic reintegration occurs at the metacognitive/metacausal level, the new situs for control, and illustrates that mind is causally nonredundant relative to brain.

Another empirical consequence of such radical transformation is illustrated in the degree of self-control exerted by such adepts. The karate master who remains poised when provoked, and so averts the fight-or-flight impulse; the yogi who melts the snow by raising his body temperature at will (Benson, 1982); the *Zen* master who slows his EEG, EKG,

EMG, GSR, and respiratory rates at will – each bears out another corollary hypothesis: The greater the metacausal complexity, the greater the functional independence from 1st-order environmental causation. Let's call this the "anti-bottom-up-determinism hypothesis". The phenomena just cited are well-documented, e.g., EEG, EKG, and, *inter alia*, GSR responses demonstrate the skilled meditator's superior degree of functional independence from stimuli.

In radical transformation, the self-structure is reconfigured so that the action-control center is no longer intimately engaged or controlled by the 1st-order stream of mental and/or environmental causation, e.g., what I presently think, feel, smell, crave, fear, etc., but is at a higher-order remove. The distance principle comes to mind. It is as if the newly-configured agent, the act-maker, is located at a new threshold, a control tower from which she looks down at a disjunctive world of functionally neutralized lower-order causal forces or equi-potent options, any of which may be toggled on-/off-line, according to autotelic designs.

This power is so rich in imagery it must be reiterated that this is the naturalized version, so let's tie these ideas down more. Recall Kim's remarks about specific supervenience relations. Let's review some here. *Lack* of attention control is linked to constrained/unfree states, an inverse implication of Chalmers' bridge law that awareness of a state enables one to verbalize it and act on it. Attentional deficiencies are at the root of many debilitating disorders such as schizophrenia, a characteristic of which is "stimulus overinclusion" – functionally, *attention overload*:

> Psychiatrists describe schizophrenics as suffering from *anhedonia*, which literally means 'lack of pleasure'. This symptom appears to be related to "stimulus overinclusion", which refers to the fact that schizophrenics are condemned to notice irrelevant stimuli, to process information whether they like it or not. (Csikszentmihalyi, 1991, p. 84)

Attention disorders confirm the anti-bottom-up-determinism and corollary metacognitive hypotheses by virtue of the pointed effects of their lack of metacognitive control.

> [A] great variety of learning disabilities have been reclassified [as] "attentional disorders", because what they have in common is lack of control over attention (*id.*).

Excessive self-consciousness, narcissism and related neurotic behaviors likely involve over-attention to any stimuli with real or perceived bearing on the self (*id.*, pp. 84-85).

Our exculpatory practices are sensitive to such illnesses that undermine autonomy. Attentional disorders involve metacausal gate-switching dysfunction – failure to regulate the on/off switching of psychoneural channels of sensory and other forms of attention. Analysis of pathology reveals what is malfunctional/functional in a system (Walter, 2002), e.g., analysis of blindness sheds light on sight. A *fully* functioning attentional system modulates the lens of sensory attention, i.e., salient world-mind input, by means of metacognitive concatenations configured by Agent's interests, and motor attention, i.e., mind-world output, by means of metacognitive concatenations configured by sensitivity to features of the environment selected by Agent's interests as alteration-worthy. Fully-functional attention autotelically modulates/integrates the quantity/quality of world-to-mind input and mind-to-world output by means of highly-differentiated metacausal configurations of Agent's will.

Our model of functional attention suggests a link between attentional skill and autotelic sophistication. Research on visual perception and cortical activation patterns supports this idea: Subjects who reported more *intrinsic motivation* in their lives – a sign of *autotelic sophistication* – required less visual focal points to reverse the figure/ground appearance of ambiguous figures.

> These findings suggest that people ... who require a great deal of outside information to form representations ... may become more dependent on ... environment for using their minds.... By contrast, people who need only a few external cues to represent events in consciousness are *more autonomous from the environment.* They have a

more flexible attention that allows them to restructure experience more easily....[123]

This confirms the anti-bottom-up-determinism and corollary metacognitive hypotheses by virtue of their metacognitive control; in stimulus over-inclusion, its privation is key. Those who report more intrinsic motivation (autotelic sophistication) engage less cortical activation to concentrate on flashes of light and sound, less than the baseline required for ordinary nonconcentration states.

> The most likely explanation [is] that the [more autotelic group] was able to reduce mental activity in every information channel but the one involved in concentrating on the flashing stimuli. This in turn suggests that [they can] screen out stimulation and to focus only on what they decide is relevant for the moment. (*id.*, p. 87-88)

These remarks are directly on point. What he says about people who have greater-than-average attentional skills is also noteworthy.

> While paying attention ordinarily involves an additional burden of information processing above the usual baseline effort, for people who have learned to control consciousness focusing attention is relatively effortless, because they can shut off all mental processes but the relevant ones. It is this flexibility of attention, which contrasts so sharply with the helpless over-inclusion of the schizophrenic, that may provide the neurological basis for the autotelic personality. (Csikszentmihalyi, 1991, p. 87-88)

These studies confirm our central and corollary hypotheses, our claims about attentional function and dysfunction, the relationship between attentional skill and autotelic sophistication, and freedom as a function of the attention-driven ability to modulate cognitive/conative input/output autotelically. *These Kim-satisfying specific-supervenience-correlations suffice for a solution to the easy problem of autonomy – self-regulation from the metacognitive attention (input) and control (output) level.*

So, the theory can provide specifications for reconfiguring dysfunctional,

---

[123] Csikszentmihalyi (1991, p. 87), emphasis added; citing Hamilton, 1976, 1981; Hamilton, Holcombe, & De la Pena, 1977; Hamilton, Haier, & Buchsbaum, 1984.

heteronomous, heterotelic systems into functional, autonomous, autotelic ones. Self-referential psychological-identification is deeply bound up with attention, the focal point of consciousness, which is plausibly why cognitive authority is sometimes cited as a *sine qua non* of freedom or personhood. By training attention, one may access the key to increased autonomy. There are other ways to train attention, but I have focused on one-pointedness and mindfulness (Goleman, 1988; Keown, 1996). Let's review them.

One-pointedness consists of concentration on an arbitrary focal point, focusing attention on a singular point and refocusing it whenever it wanders; mindfulness consists of non-conceptual awareness of the contents of consciousness, simply paying attention to whatever objects/states enter/exit consciousness. One-pointedness develops concentration, the ability to control the gaze of consciousness, and mindfulness develops detachment, the ability to dissociate from the contents of consciousness, and insight, the ability to see things as they are (Goleman, 1988). That these skills may increase metacausal functioning seems implausible, for research has shown that what the mind pays attention to is determined:

> Essentially, long-term memory ... becomes the [sensory] filter, deciding what to block from short-term storage (and [thus] awareness), thereby determining indirectly what to accept for eventual storage in long-term memory itself. (Erdelyi, 1974, p. 19)

On this model, memory scans sensory information at every stage of its processing and filters it for salience *before* it enters awareness; only a fraction available at any moment enters consciousness (Goleman, 1986). But while LTM is the key determinant in filtering at the "doors of perception" (Huxley, 1954), attention may be consciously selective, influence LTM, and modify the sensory filter: We can scan for something, so awareness can modify how the filter operates – through LTM, the activity of which is not evident to awareness (Goleman, 1986). But the self-system determines what LTM marks as salient – that the determination of the filter by awareness is indirect is irrelevant: We are never directly aware

of, say, the neural bases of any elements of our mental lives, but so what? Insofar as one-pointedness training is one way to voluntarily control attention, and attention determines what gets filtered perceptually, it is a way to regulate 1st-order cognitive – *environmental* – input. Walter's (2002) ranking of environmentally-triggered psychoneural states at the bottom of the hierarchy coheres well with the distance principle.

The other meditative skill, mindfulness, can be used to regulate whatever passes through the first meditative filter: If one is skillful at merely *attending to* or *noting* objects or qualities of consciousness, one is skillful at not having to act on or react to them. This can liberate one from compulsive, neurotic, and other unfree behaviors; e.g., the *dissociative distancing* that attends mental states characterized as being aware *that anxiety is arising*, *that powerful hunger urges are arising*, or *that strong feelings of loneliness are arising* differs radically from that of mental states characterized as *identifying with* the thoughts represented in the sentences, "I should take a sedative", "I'm going to devour that Chinese food", or "I've got to speak to someone". When one is in the grip of those states of mind, one is typically unconscious that one is in their grip.

The difference is not at the propositional level, but at the level of *skill* in maintaining mental states characterized as *observing* mental contents as opposed to *being engaged* with them, a skill developed by the practice of mindfulness (Keown, 1996). The difference is also a matter of the degree to which one is on-line with respect to the contents of consciousness, and is confirmed by many generations of practitioners. Theological lore is data for a theory of will; it contains the most complete record on the phenomenology of *akrasia*, 'spirit versus flesh', sin, etc. (Ainslie, 2001). Thus, the contemplative virtues of long-term meditators are the subject of platitudes in contemplative communities, e.g., equanimity; they function as litmus tests of progress in contemplative communities throughout history. Buddhist

psychology classifies mental states or behavioral characteristics (e.g., lust, forgiveness) as virtuous/skillful or vicious/unskillful to the extent that they tend to further or hinder the development of metacognitive control, measured as attainment of advanced meditative states (Goleman, 1988) – a naturalistic approach to the virtues our theory may happily adopt.

A key virtue-skill is restraint. One is free from compulsive behavior if he can exercise restraint over his desires, if he can go off-line to eschew acting on them. Attention training strengthens this ability. Insofar as these attention-training techniques develop our top-down ability to regulate 1st-order, on-line, S/R-causation from a higher-order attentional vantage, they provide a generic means of redesigning our agential structures – the figurative keys to the gates in the metacausal kingdom. Since the more complex, the more fully formed the person, it follows that the formulaic advice on how to redesign our metacausal structures constitutes a way to decrease wantonness and increase personhood.

If attention-training techniques help reconfigure our metacausal structures, then we have the beginnings of prescriptions not only for increased agency, but also for educational and moral psychology. Recall what William James had to say about attention and learning:

> The faculty of voluntarily bringing back a wandering attention over and over again is the very root of judgment, character, and will.... An education that should improve this faculty would be the education *par excellence*. (1910, quoted in Goleman, 1991, p. 98)

Meditation is the elixir James seeks. It has obvious implications for psychology and ethics.

Dennett claimed (1984) there will never be 'a' solution to the free will problem because there are many such problems. We examined many dimensions of the free will problem, and exposed the heterogeneity and ubiquity of the data, of the phenomenology, and of the intuitions of our daily experience of mundane autonomy, which we may now call "non-phi-fi autonomy". We also supported the claim that most of these data and their

associated phenomenology are misrepresented by single-criterion conceptions of autonomy, responsibility, and related notions specifically designed to be what may be called 'phi-fi-resistant', akin to the way epistemic-competency theories are designed to be Gettier-resistant.

To gain an integrated perspective on this unwieldy collection of pre-philosophical and philosophical problems, we appealed to Russell (2002), who with intellectual economy groups these problems in three classes of pessimism: close-, middle, and horizon-level. We replied to horizon-range pessimism in our critique of CON, and to close-/middle pessimisms with the cognitive architectural blueprints for easy autonomy, a metacausal theory of will and of autonomy. Our theory circumvents horizon-level pessimism about ultimate control and provides naturalized notions of *possible, can*, and *free*, opening a door for replies to close- and middle-level pessimisms in advance of positive arguments. We provided an empirically-supported just-so story for, and a causal/functional analysis of, the powers involved in both hierarchical and nonhierarchical wills. We developed a theory of will therewith, identified the specifications for easy-autonomy, and showed how our theory satisfies them. We specified the metacausal principle of autonomy, showed how it handles phi-fi-Frankfurt-cases, the moral asymmetry problem, and *akrasia*, how it grounds a naturalized form of F&R's regulative-control, and how it supports a model of self-formation.

Our theory has plausible rebuttals for all three pessimisms, and grounds our reactive attitude, moral, and other normative attributive practices. It also promises to unite libertarian, Kantian, and many related inflated intuitions about mind/action under a natural conception of self that squares with neurophilosophy (Walter, 2002) and common sense, and has pragmatic implications for education, psychology, and ethics.

Since our solution to the easy problem is plausible, it renders hard problems otiose by way of its parsimony and its other explanatory virtues. These include its coherence with

what we know in evolutionary science, developmental psychology, behavioral science, and neurophilosophy; its reflective equilibrium with the relevant pre-reflective data, phenomenology, and intuitions; and its ability to absorb, deflate, naturalize, and otherwise account for the intuitions – and even save the reflective insights – of its competitors.

Further empirical studies may also be expected to supply the neural details for the causal/functional specifications of our solution. In light of its empirical base, improvements over competing theories, and philosophical and practical applications, our theory covers much ground. While there may be other solutions, to the extent that our theory collects and responds to all forms of pessimism under one all-encompassing explanatory blanket capable of absorbing all the surviving intuitions of all its previous competitors, our theory comes close to rebutting Dennett's negative claim about 'a' solution. If this is plausible, it is something worth knowing, a worthwhile project in its own right.

References

Adams, Robert Merrihew (1991). "An Anti-Molinist Argument", *Philosophical Perspectives* 5:353-354.

Adler, Jonathan (1995). "Free Will". Lect.: CUNY Graduate School and University Center (Spring 1995).

Adler, Jonathan (1997). "The Ethics of Belief". Lect.: CUNY Graduate School and University Center (Fall 1997).

Adler, Jonathan (2002). *Belief's Own Ethics*. Cambridge: MIT Press.

Ainslie, George (2001). *Breakdown of Will*. Cambridge: Cambridge UP..

Anand, B.K., C.S. Chhina, and Baldev Singh (1961). "Some Aspects of Electroencephalographic Studies in Yogis", *Electroencephalography and Clinical Neurophysiology* 13:452-56.

Andersen, Peter Bogh, et al., eds. (2000). *Downward Causation*. Aarhus, Denmark: Aarhus UP.

Aristotle (1925). "Continence and incontinence", *Nichomachean Ethics*, book vii, chapters 1-3, translated by W.D. Ross in *The Oxford Translation of Aristotle*, ed. W.D. Ross (1925) vol. IX, reprinted in Mortimore (1971), pp. 63-68.

Aristotle (1941). "Fatalism, Voluntary Action, and Choice", *On Interpretation*, in *The Basic Works of Aristotle*, ed. Richard McKeon (New York: Random House, 1941), pp. 45-48, reprinted in Hoy & Oaklander (1991), pp. 300-302.

Atmanspacher, Harald, and Eva Ruhnau, eds. (1997). *Time, Temporality, Now: Experiencing Time and Concepts of Time in an Interdisciplinary Perspective*. Berlin: Springer-Verlag.

Austin, James H. (1998). *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. Cambridge: MIT Press.

Ayer, A.J. (1954). "Freedom and Necessity", *Philosophical Essays*. London: Macmillan.

Baron-Cohen, Simon (1995). *Mindblindness*. Cambridge: MIT Press.

Baron-Cohen, S., A.M. Leslie, and U. Frith (1995). "Does the Autistic Child Have a Theory of Mind?", *Cognition* 21:37-46.

Baron-Cohen, S. and Cross, P. (1992). "Reading the Eyes: Evidence for the Role of Perception in the Development of a Theory of Mind", *Mind and Language* 7:172-86.

Beckermann, Ansgar, Hans Flohr, and Jaegwon Kim, eds. (1992). *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism.* Berlin: Walter de Gruyter.

Benson, Herbert, M.D. (1975). *The Relaxation Response* (New York: Morrow).

Benson, Herbert, M.D. (1982). "Body Temperature Changes During the Practice of *gTum-mo* yoga", *Nature* 298, pp. 234-6.

Benson, Herbert, M.D. (1996). *Timeless Healing: The Power and Biology of Belief.* New York: Scribner.

Benson, H., Malhotra, M.S., *et al.* (1990). "Three Case Reports of the Metabolic and Electroencephalographic Changes During Advanced Buddhist Meditative Techniques", *Behavioral Medicine* 16, pp. 90-5.

Bernstein, Mark (2002). "Fatalism" in Kane (2002a).

Berofsky, Bernard (2002). "Ifs, Cans, and Free Will: The Issues" in Kane (2002a).

Bishop, Robert C. (2002). "Chaos, Indeterminism, and Free Will" in Kane (2002a).

Blackburn, Simon (1992). "Theory, Observation and Drama", *Mind and Language* 7:187-203.

Blackburn, Simon (1998). *Ruling Passions: A Theory of Practical Reasoning.* New York: Oxford UP.

Blackmore, Susan (1999). *The Meme Machine.* New York: Oxford UP.

Blakeslee, Sandra (1997). "Recipe for a Brain: Cups of Genes and a Dash of Experience?", *Science Times, The New York Times*, p. F4 (November 4, 1997).

Block, Ned (1997). "Anti-Reductionism Slaps Back", *Philosophical Perspectives* 11(5):107-31.

Block, N., O. Flanagan, and G. Güzeldere, eds. (1997). *The Nature of Consciousness: Philosophical Debates.* Cambridge: MIT Press.

Bogdan, Radu J. (1997). *Interpreting Minds.* Cambridge, MA: MIT Press.

Bogdan, Radu J. (2003). *Minding Minds: Evolving a Reflexive Mind by Interpreting Others*, Cambridge: MIT Press.

Bond, Edward J. (1983). *Reason and Value.* New York: Cambridge UP.

Bok, Hilary (1998). *Freedom and Responsibility.* Princeton: Princeton UP.

Bowlby, J. (1982). *Attachment.* New York: Basic Books.

Broad, C.D. (1925). *Mind and Its Place in Nature.* London: Routledge & Kegan Paul.

Butterworth, G. (1991). "The Ontogeny and Phylogeny of Joint Visual Attention" in Whiten (1991).

Butterworth, G. (1995). "Origins of Mind in Perception and Action" in Moore and Dunham (1995).

Byrne, R.W. and A. Whiten, eds. (1991). "Computation and Mindreading in Primate Tactical Deception" in Whiten (1991).

Byrne, R.W. (1995). "The Ape Legacy: The Evolution of Machiavellian Intelligence and Anticipatory Interactive Planning", in Goody (1995).

Campbell, C. (1967). *In Defense of Free Will.* London: Allen & Unwin.

Carruthers, P. And P.K. Smith, eds. (1996). *Theories of Theories of Mind.* Cambridge: Cambridge UP.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford UP.

Chalmers, David (1997). "The Puzzle of Conscious Experience", *Scientific American* (December 1995); reprinted in *Scientific American Mysteries of the Mind*, Special Issue 7(1):30-7 (1997).

Changeau, J.-P. and S. Dehaene (1989). "Neuronal Modes of Cognitive Functions," *Cognition* 33:63-109.

Changeau, J.-P. and S. Dehaene (1995). "Neuronal Models of Cognitive Functions Associated with the Prefrontal Cortex," in Damasio et al. (1996), pp. 125-44.

Chisolm, Roderick (1982). "Human Freedom and the Self", in Watson (1982), pp. 24-35.

Collins, Arthur (1984). "Action, Causality, and Teleological Explanation", *Midwest Studies in Philosophy* vol. IX. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Conniff, Richard (2004). "Reading Faces", *Smithsonian* (January 2004).

Cornford, Francis M., trans. (1941). *The Republic of Plato.* London: Oxford UP.

Crick, Francis (1994). *The Astonishing Hypothesis.* New York: Scribner's & Sons.

Crick, F. and C. Koch (1997). "The Problem of Consciousness", *Scientific American* (September 1992); reprinted in *Scientific American Mysteries of the Mind*, Special Issue 7(1):19-26 (1997).

Csikszentmihalyi, Mihalyi (1991). *Flow: The Psychology of Optimal Experience*. New York: Harper Perennial.

Cummins, Robert (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Damasio, A.R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Damasio, A.R., H. Damasio, and Y. Christen, eds. (1996). *Neurobiology of Decision Making*. New York: Springer.

Davidson, Donald (1968). "Actions, Reasons, and Causes", *The Philosophy of Action*. Oxford: Oxford UP.

Davidson, Donald (1980). "How Is Weakness of Will Possible?", *Essays on Actions and Events*. Oxford: Clarendon Press.

Dawkins, Richard (1976). *The Selfish Gene*. Oxford: Oxford UP.

Dennett, Daniel (1981). "On Giving Libertarians What They Say They Want", *Brainstorms* (Cambridge: MIT Press).

Dennett, Daniel (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: MIT Press.

Dennett, Daniel (2003). *Freedom Evolves*. New York: Viking.

De Waal, F. (1982). *Chimpanzee Politics*. Baltimore: Johns Hopkins UP.

Donnellan, Keith (1975). "Reference and Definite Descriptions", *Philosophical Review* 77:281-304.

Double, Richard (2002). "Metaethics, Metaphilosophy, and Free Will Subjectivism" in Kane (2002a).

Dretsky, Fred (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dunn, J. (1988). *The Beginnings of Social Understanding*. Oxford: Blackwell.

Dworkin, Ronald (1970). "Acting Freely", *Nous* 4:367-83.

Ekstrom, Laura Waddell (2000). *Free Will: A Philosophical Study*. Boulder: Westview Press.

Ekstrom, Laura Waddell (2002). "Libertarianism and Frankfurt-style Cases" in Kane (2002a).

Elster, John (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge UP.

Elster, John (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge UP.

Engel, Andreas K., Pieter R., Roelfsema, Peter Konig, and Wolf Singer (1997). "Neurophsyiological Relevance of Time", in Atmanspacher and Ruhnau (1997), pp.133-57.

Erdelyi, M. (1974). "A New Look at the New Look: Perceptual Defense and Vigilance," *Psychological Review, 81,* 1-25.

Fischer, John Martin (1987). "Responsiveness and Moral Responsibility" in Schoeman (1987); reprinted in Pereboom (1997), pp. 214-41.

Fischer, John Martin (1994). *The Metaphysics of Free Will*. Cambridge: Blackwell.

Fischer, John Martin (2002). "Frankfurt-type Examples and Semi-Compatibilism" in Kane (2002a).

Fischer, John Martin, and Mark Ravizza, S.J. (1993). *Perspectives on Moral Responsibility*. Ithaca: Cornell UP.

Fischer, John Martin, and Mark Ravizza, S.J. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge UP.

Flanagan, Owen (1992). *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Fodor, Jerry (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: Bradford-MIT.

Fodor, Jerry (1992). "Philosophy of Mind: Recent Topics". Lect.: CUNY Graduate School (Spring 1992).

Forrest-Presley, D.L., G.E. Mackinnon, and T. Gary Waller, eds. (1985). *Metacognition, Cognition, and Human Performance*. Orlando: Academic Press.

Frankfurt, Harry (1969). "Alternate Possibilities and Moral Responsibility", *Journal of Philosophy* 66 (December 1969):23; reprinted in Frankfurt (1988).

Frankfurt, Harry (1971). "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68 (January 1971) 1:5-20, reprinted in Frankfurt (1988).

Frankfurt, Harry (1977). "Identification and Externality", in Rorty (1977), 239-51, reprinted in Frankfurt (1988).

Frankfurt, Harry (1987). "Identification and Wholeheartedness", in Schoeman (1987), reprinted in Frankfurt (1988).

Frankfurt, Harry (1988). *The Importance of What We Care About.* Cambridge: Cambridge UP.

Frankfurt, Harry (1999). *Necessity, Volition and Love.* Cambridge: Cambridge UP.

Gibson, K.R., and T. Ingold, eds. (1993). *Tools, Language, and Cognition in Human Evolution.* Cambridge: Cambridge UP.

Gilbert, Scott F., and Sahotra Sartar (2000). "Embracing Complexity: Organicism for the Twenty-first Century", *Developmental Dynamics* 219:1-9.

Ginet, Carl (2002). "Reasons Explanations of Action: Causalist versus Noncausalist Accounts" in Kane (2002a).

Goldman, Alvin I. (1970). *A Theory of Human Action.* Englewood Cliffs, NJ: Prentice-Hall.

Goldman, Alvin I. (1992). "In Defense of the Simulation Theory", *Mind and Language* 7:104-19.

Goldman, Alvin I. (1993). "Ethics and Cognitive Science", *Ethics*, Vol. 103:2, pp. 337-60.

Goleman, Daniel (1986). *Vital Lies, Simple Truths: The Psychology of Self-Deception.* New York : Simon & Schuster.

Goleman, Daniel (1988). *Varieties of the Meditative Experience.* New York: E.P. Dutton.

Goleman, Daniel (1991). "Tibetan and Western Models of Mental Health" in Goleman and Thurman (1991), pp. 89-102.

Goleman, Daniel (1995). *Emotional Intelligence: Why It Can Matter More Than IQ.* New York : Bantam Books.

Goleman, Daniel, narr. (2003). *Destructive Emotions: How Can We Overcome Them? A Scientific Dialogue with the Dalai Lama.* New York: Bantam Books.

Goleman, Daniel and Thurman, Robert A.F., eds. (1991). *Mind Science: An East-West Dialogue.* Boston: Wisdom Publications.

Gomez, J.C. (1990a). "Primate Tactical Deception and Sensorimotor Social Intelligence", *Behavioral and Brain Sciences* 13:414-15.

Gomez, J.C. (1990b). "The Emergence of Intentional Communication as a Problem-Solving Strategy in the Gorilla", in Parker and Gibson (1990).

Gomez, J.C. (1991). "Visual Behavior as a Window for Reading the Mind of Others in Primates", in Whiten (1991).

Gomez, J.C. (1994). "Mutual Awareness in Primate Communication: A Gricean Approach", in Parker *et al.* (1994).

Gomez, J.C. (1996a). "Second Person Intentional Relations and the Evolution of Social Understanding", *Behavioral and Brain Sciences* 19:129-30.

Gomez, J.C. (1996b). "Ostensive Behavior in Great Apes: The Role of Eye Contact", in Russon *et al.* (1996).

Goody, E.N., ed. (1995). *Social Intelligence and Interaction.* Cambridge: Cambridge UP.

Gopnik, Alison and Wellman, Henry M. (1992). "Why the Child's Theory of Mind Really *Is* a Theory", *Mind and Language* 7:145-71.

Gordon, Robert (1992a). "The Simulation Theory: Objections and Misconceptions", *Mind and Language* 7:11-34.

Gordon, Robert (1992b). "Reply to Stich and Nichols", *Mind and Language* 7:87-97.

Gordon, Robert (1992c). "Reply to Perner and Howes", *Mind and Language* 7:98-103.

Greenwood, John (1999). "Simulation, Theory-Theory and Cognitive Penetration: No 'Instance of the Fingerpost'", *Mind and Language* 14(1):32-56.

Grice, G.R. (1980). "Are There Reasons for Acting?", *Midwest Studies in Philosophy* vol. III. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Güven Güzeldere (1997). "Is Consciousness the Perception of What Passes in One's Own Mind?", *Conscious Experience*, ed. T. Metzinger. Paterborn: Schoeningh-Verlag (1995); reprinted in Block, Flanagan, and Güzeldere (1997).

Hacking, Ian (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science* (Cambridge: Cambridge UP).

Hamilton, J.A. (1976). "Attention and Intrinsic Rewards in the Control of Psychophysiological States", *Psychotherapy and Psychosomatics* 27:54-61.

Hamilton, J.A. (1981). "Attention, Personality, and Self-Regulation of Mood: Absorbing Interest and Boredom", in Maher (1981).

Hamilton, Haier, and Buchsbaum (1984). "Intrinsic Enjoyment and Boredom Coping Scales: Validation with Personality Evoked Potential and Attentional Measures", *Personality and Individual Differences* 5(2):183-93.

Hamilton, Holcomb, and De la Pena (1977). "Selective Attention and Eye Movements while Viewing Reversible Figures", *Perceptual and Motor Skills* 44:639-44.

Hampshire, Stuart (1975). *Freedom of the Individual.* Princeton: Princeton UP.

Hampshire, Stuart (1987). *Spinoza: An Introduction to His Philosophical Thought.* New York: Penguin Books.

Harris, Paul L. (1992). "From Simulation to Folk Psychology: The Case for Development", *Mind and Language* 7:120-44.

Hodgson, David (2002). "Quantum Physics, Consciousness, and Free Will" in Kane (2002a).

Honderich, Ted (1993). *How Free Are You?* New York: Oxford UP.

Honderich, Ted (2002). "Determinism as True, Compatibilism and Incompatibilism as False, and the Real Problem" in Kane (2002a).

Hoy, Ronald C. And Oaklander, Nathan, eds. (1991). *Metaphysics: Classic & Contemporary Readings.* Belmont: Wadsworth.

Hume, David (1748). "On Liberty and Necessity", *An Enquiry Concerning Human Understanding*, Sec. 8, reprinted in Hoy and Oaklander (1991), pp. 328-337.

Humphrey, Nicholas (1982). "Consciousness: a Just So Story", *New Scientist* 95.

Humphreys, Paul (1997). "How Properties Emerge", *Philosophy of Science* 64:1-17 (March 1997).

Huxley, Aldous (1954). *The Doors of Perception.* New York: Harper.

James, William (1910). *The Principles of Psychology.* New York: Dover.

Kane, Robert, ed. (2002a). *The Oxford Handbook of Free Will.* Oxford: Oxford UP.

Kane, Robert (2002b). "Some Neglected Pathways in the Free Will Labyrinth", in Kane (2002a).

Kapitan, Tomis (2002). "A Master Argument for Incompatibilism" in Kane (2002a).

Katz, Jerrold (1997). "The Problem in 20th Century Philosophy". Lect.: CUNY Graduate School and University Center (Fall 1997).

Kauffman, Stuart (1995). *At Home in the Universe*. New York: Oxford UP.

Keown, Damien (1996). *Buddhism: A Very Short Introduction*, Oxford: Oxford UP.

Kenny, Anthony, *et al.* (1972). *The Nature of Mind*. Edinburgh: University Press.

Kim, Jaegwon (1979). "Causality, Identity, and Supervenience in the Mind-Body Problem", *Midwest Studies in Philosophy* vol. IV. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Kim, Jaegwon (1984). "Epiphenomenalism and Supervenient Causation", *Midwest Studies in Philosophy*, vol. IX. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Kim, Jaegwon (1989). "Supervenience as a Philosophical Concept". Third Metaphilosophy Address. CUNY Graduate School and University Center (May 1989).

Kim, Jaegwon (1992). "Multiple Realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research* 52:1 (March 1992).

Kim, Jaegwon (1999). "Making Sense of Emergence", *Philosophical Studies* 95:3-36.

Kim, Jaegwon (2000). "Making Sense of Downward Causation", in Andersen et al. (2000), pp. 305-321.

Kinsbourne, Marcel (1992). "Integrated Field Theory of Consciousness", in Marcel and Bisiach (1992).

Kitcher, Philip (1987). "*Apriority* and Necessity" in Moser (1987), pp. 190-207.

Kripke, Saul (1972). *Naming and Necessity*. Cambridge: Harvard UP.

Kuppers, Bernd-Olaf (1992). "Understanding Complexity", in Beckermann, Flohr, and Kim (1992), pp. 241-56.

Landesman, Charles (2003). Email correspondence.

Langer, Susanne K. (1967). *Mind: An Essay on Human Feeling*. Baltimore, MD: The John Hopkins UP.

LeDoux, Joseph (1997). "Emotion, Memory and the Brain", *Scientific American* (June 1994); reprinted in *Scientific American Mysteries of the Mind*, Special Issue 7(1):68-75 (1997).

Lehrer, Keith (1966). *Freedom and Determinism.* New York: Random House.

Lehrer, Keith (1980). "Preferences, Conditionals, and Freedom", in van Inwagen (1980), pp. 76-96.

Levin, Michael (1979). *Metaphysics and the Mind-Body Problem.* Oxford: Oxford UP.

Levin, Michael (2003). "Free Will and Moral Responsibility", in Pojman (2003).

Lewis, David (1979). "Counterfactual Dependence and Time's Arrow", *Nous*13:455-76; reprinted in Lewis (1986a).

Lewis, David (1986a). *Philosophical Papers*, Vol. II. Oxford: Oxford UP.

Lewis, David (1986b). "Are We Free To Break the Laws?", *Theoria* 47 (1981), pp 113-21; reprinted in Lewis (1986a), pp. 291-8.

Lewis, David (2000). "Causation as Influence", *Journal of Philosophy* 97:182-97.

Libet, Benjamin (2002). "Do We Have Free Will?" in Kane (2002a).

Locke, John (1690). *An Essay Concerning Human Understanding* (A.C. Fraser edition). New York: Dover, 1959.

Madow, Leo and Snow, Laurence H., eds. (1970). *The Psychodynamic Implications of Physiological Studies on Sensory Deprivation.* Springfield, Ill.: Thomas.

Maher, B.A. ed. (1981). *Progress in Experimental Personality Research* 10:282-315.

Marcel, A.J. and E. Bisiach, eds. (1992). *Consciousness in Contemporary Science*, Oxford: Oxford UP.

Martin, C.B. (1994). "Dispositions and Conditionals", *The Philosophical Quarterly* 44(174):1-8.

McGuigan, Frank J. (1978). *Cognitive Psychophysiology: Principles of Covert Behavior.* Englewood Cliffs, NJ: Prentice-Hall.

McGinn, Colin (2003). "Screen Dreams". Sprague-Taylor Lecture Series. Brooklyn College of the City University of New York (Spring 2003).

Mele, Alfred R. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control.* New York: Oxford UP.

Mele, Alfred R. (1995). *Autonomous Agents: From Self-Control to Autonomy*, New York: Oxford UP.

Mele, Alfred R. (2002). "Autonomy, Self-Control, and Weakness of Will" in Kane (2002a).

Mele, Alfred R. (2003). "Agnostic Autonomism", www.ucl.ac.uk/~uctytho/dfwVariousMele. html.

Mele, Alfred R., and David Robb (1998). "Rescuing Frankfurt-style Cases." *The Philosophical Review* 107:97-112.

Melzack, Ronald (1997). "Phantom Limbs", *Scientific American* (April 1992); reprinted in *Scientific American Mysteries of the Mind*, Special Issue 7(1):84-91 (1997).

Metcalfe, Janet and Arthur P. Shimamura, eds. (1994). *Metacognition: Knowing about Knowing*. Cambridge: MIT Press.

Milgram, Stanley (1963). "Behavioral Study of Obedience", *Journal of Abnormal and Social Psychology*, 67, 371-378.

Moore, C. And P. Dunham, eds. (1995). *Joint Attention*. Hillsdale, NJ: Erlbaum.

Mortimore, G.W., ed. (1971). *Weakness of Will*. London: MacMillan.

Moser, Paul K., ed. (1987). *A Priori Knowledge*. New York: Oxford UP.

Nagel, Thomas (1970). *The Possibility of Altruism*. Oxford: Clarendon Press.

Neely, Wright (1974). "Freedom and Desire", *Philosophical Review* 83:32-54.

Neisser, U., ed. (1993). The Perceived Self. Cambridge: Cambridge UP.

Nelson, Thomas O., ed. (1992). *Metacognition: Core Readings*. Boston: Allyn and Bacon.

Nisbett, Richard and Ross, Lee (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Northoff, G. (1999). "Psychomotor Phenomena As Paradigmatic Examples of Functional Brain Organisation and the Mind-Brain Relationship", *Philosophy, Psychology, and Psychiatry* 6(3):199-230.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge: Harvard UP.

O'Connor, Timothy (1995). *Agents, Causes, & Events: Essays on Indeterminism and Free Will*. New York: Oxford UP.

O'Connor, Timothy (2002). "Libertarian Views: Dualist and Agent-Causal Theories" in Kane (2002a).

Olton, David S. and Aaron R. Noonberg (1980). *Biofeedback: Clinical Applications in Behavioral Medicine*. Englewood Cliffs: Prentice-Hall.

Parfit, Derek (1986). *Reasons and Persons*. New York: Oxford University Press.

Parker, S.T., and K.R. Gibson, eds. (1990). *"Language" and Intelligence in Monkeys and Apes*. Cambridge: Cambridge UP.

Parker, S.T., *et al.*, eds. (1994). *Self-Awareness in Animals and Humans*. Cambridge: Cambridge UP.

Patanjali (1953). *Yoga Sutras*. Trans. Swami Prabhavananda and Christopher Isherwood as *How To Know God: The Yoga Aphorisms of Patanjali*. Hollywood, CA: Vedanta Press.

Pears, David (1984). *Motivated Irrationality*. Oxford: Clarendon Press.

Pereboom, Derk, ed. (1997). *Free Will*. Indianapolis: Hackett Publishing Co.

Pereboom, Derk (2002). "Living without Free Will: The Case for Hard Incompatibilism" in Kane (2002a).

Perls, Fritz (1947). *Ego, Hunger, and Aggression*. New York: Vintage.

Perls, Fritz, R.F. Hefferline, and P. Goodman (1951). *Gestalt Therapy*. New York: Julian.

Perner, Josef and Howes, Deborrah (1992). "'He Thinks He Knows': And More Developmental Evidence Against the Simulation (Role Taking) Theory", *Mind and Language* 7:72-86.

Petrosky, T., and I. Prigogine (1997). "The Extension of Classical Dynamics for Unstable Hamiltonian Systems", *Computers & Mathematics with Applications* 34:1-44.

Pinker, Steven (1997). *How the Mind Works*. New York: W.W. Norton & Co.

Plato (1941). *The Republic of Plato*. Trans. F.M. Cornford. Oxford: Oxford UP.

Plato (1952). *Plato's Phaedrus*. Trans. R. Hackford. Cambridge: Cambridge UP.

Pojman, Louis, ed. (2003). *Introduction to Philosophy*, 2nd ed. New York: Oxford UP.

Povinelli, D.J. (1996). "Chimpanzee Theory of Mind?" in Carruthers and Smith (1996).

Povinelli, D.J., and T.J. Eddy (1996). *What Young Chimpanzees Know about Seeing*. Monographs of the Society for Research in Child Development, vol. 61, no. 3. Chicago: University of Chicago Press.

Price, H.H. (1953). "Universals and Resemblances", *Thinking and Experience*. Cambridge: Harvard UP.

Quine, W.V.O. (1953). *From a Logical Point of View*. New York: Harper.

Reynolds, P.C. (1993). "The Complementation Theory of Language and Tool Use" in Gibson and Ingold (1993).

Robbins, Jim (1997). "Biofeedback Offers Help to Hyperactive Children", *Science Times, The New York Times*, p. F7 (November 11, 1997).

Rorty, Amelie O. (1977). *The Identity of Persons*. Berkeley: University of California Press.

Rosenthal, David (2000). "Consciousness and Metacognition", manuscript, reprinted in Sperber (2000).

Rosenthal, David (1997). "A Theory of Consciousness", ZiF Technical Report. Bielefeld, Germany; reprinted in Block, Flanagan, and Güzeldere, eds. (1997).

Rosenthal, David, ed. (1991a). *The Nature of Mind*. New York: Oxford UP.

Rosenthal, David (1991b). "Two Concepts of Consciousness", *Philosophical Studies* 94(3):329-59 (May 1986); reprinted in Rosenthal (1991a).

Russell, Paul (2002). "Pessimists, Pollyanas, and the New Compatibilism" in Kane (2002a).

Russon, A.E., *et al.*, eds. (1996). *Reaching into Thought*. Cambridge: Cambridge UP.

Sapolsky, Robert (1997). "Testosterone Rules", *Discover* 18(3):44-50 (March 1997).

Schiffer, Steve (1987). *Remnants of Meaning*. Cambridge, MA: MIT Press.

Schiffer, Steve (1990). "Physicalism", *Philosophical Perspectives* 4, *Action Theory and Philosophy of Mind*, pp.153-185.

Schiffer, Steve (1991). "*Ceteris Paribus* Laws", *Mind* 100(1):1-17 (January 1991).

Schiffer, Steve (1992). "The Naturalization of Content". Lect.: CUNY Graduate School (Fall 1992).

Schoeman, Ferdinand, ed. (1987). *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. New York: Cambridge UP.

Schultz, Duane P. (1965). *Sensory Restriction: Effects on Behavior*. New York: Academic Press.

Schwartz, Gary E. and Shapiro, David (1976). *Consciousness and Self-Regulation*. New York: Plenum Press.

Searle, John R. (1984). "The Freedom of the Will", *Minds, Brains and Science*. Cambridge: Harvard UP.

Searle, John R. (1992). *The Rediscovery of the Mind*. Cambridge: MIT Press.

Searle, John R. (1994). "Animal Minds", *Midwest Studies in Philosophy*, vol. XIX. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Shatz, David (1986). "Free Will and the Structure of Motivation", *Midwest Studies in Philosophy* 10:451-482. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Singer, Peter, ed. (1994). *Ethics*. Oxford: Oxford UP.

Slote, Michael (1980). "Understanding Free Will", *Journal of Philosophy* 77:136-51.

Slote, Michael (1982). "Selective Necessity and the Free-Will Problem", *Journal of Philosophy* 79:5-24.

Smilansky, Saul (2002). "Free Will, Fundamental Dualism, and the Centrality of Illusion" in Kane (2002a).

Sober, Elliot (1995). *Core Questions in Philosophy* 2d. Englewood Cliffs, NJ: Prentice-Hall.

Sorensen, Roy (1992). *Thought Experiments*. New York: Oxford UP.

Sperber, Dan, ed. (2000). *Metarepresentation: Proceedings of the Tenth Vancouver Cognitive Science Conference*. New York: Oxford UP.

Stalnaker, Robert C. (1984). *Inquiry*. Cambridge, MA: MIT Press..

Stelmach, George E., ed. (1976). *Motor Control: Issues and Trends*. New York: Academy Press.

Stich, Stephen and Nichols, Shaun (1992). "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language*7:35-71.

Stich, Stephen and Nichols, Shaun (1993). "Second Thoughts on Simulation Theory", *RuCCS TR-11*. Technical Report #11, Rutgers University Center for Cognitive Science.

Stout, G.F. (1931). *Mind & Matter*. Cambridge: The University Press.

Strawson, Galen (1986). *Freedom and Belief.* Oxford: Clarendon Press.

Strawson, Galen (1994). *Mental Reality.* Cambridge: MIT Press.

Strawson, Galen (1997). "On Being a Materialist: Realistic Monism". Lect.: CUNY Graduate School and University Center (Fall 1997).

Strawson, Galen (2002). "The Bounds of Freedom" in Kane (2002a).

Strawson, Peter F. (1962). "Freedom and Resentment", *Proceedings of the British Academy* 48:1-25, reprinted in Watson (1982), pp. 59-80.

Stump, Eleonore (1993). "Sanctification, Hardening of the Heart, and Frankfurt's Concept of the Will", in Fisher and Ravizza (1993), pp. 211-34.

Taylor, Christopher and Dennett, Daniel (2002). "Who's Afraid of Determinism? Rethinking Causes and Possibilities" in Kane (2002a).

Taylor, Paul W. (1986). *Respect for Nature: A Theory of Environmental Ethics.* Princeton: Princeton UP.

Taylor, Richard (1983). "Fate", *Metaphysics.* New York: Prentice-Hall, pp. 51-62.

Tomasello, M. And J. Call (1997). *Primate Cognition.* New York: Oxford UP.

Trevarthen, C. (1993). "The Self Born in Intersubjectivity", in Neisser (1993).

Umlità, Carlo (1992). "The Control Operations of Consciousness", in Marcel and Bisiach (1992).

Van Inwagen, Peter (1975). "The Incompatibility of Free Will and Determinism", *Philosophical Studies* 27:185-99, reprinted in Watson (1982).

Van Inwagen, Peter (1978). "Ability and Responsibility", *The Philosophical Review* 87:201-24.

Van Inwagen, Peter, ed. (1980). *Time and Cause.* Dordrecht: Reidel.

Van Inwagen, Peter (1983). *An Essay on Free Will.* Oxford: Clarendon Press.

Van Inwagen, Peter (2002). "Free Will Remains a Mystery" in Kane (2002a).

Velleman, David (1992). "What Happens When Someone Acts?", *Mind* 101:461-81.

Vendler, Zeno (1984). "Agency and Causation", *Midwest Studies in Philosophy*, vol. IX. French, Uehling, and Wettstein, eds. Minneapolis: University of Minneapolis Press.

Villanueva, Enrique (1991). *Consciousness*. Atascadero, CA: Ridgeview Publishing Co.

Walter, Henrik (2001). *Neurophilosophy of Free Will*. Cambridge, MA: MIT Press.

Walter, Henrik (2002). "Neurophilosophy of Free Will" in Kane (2002a).

Wason, P. And P. Johnson-Laird (1972). *Psychology of Reason: Structure and Content*. Cambridge: Harvard UP.

Watson, Gary (1975). "Free Agency", *Journal of Philosophy* 72(8):205-20.

Watson, Gary, ed. (1982). *Free Will*. Oxford: Oxford UP.

Weinert, Franz E. and Rainer H. Kluwe, eds. (1987). *Metacognition, Motivation, and Understanding*. Hillside, NJ: Lawrence Erlbaum Associates.

Weiskrantz, Lawrence (1992). "Neuropsychology of Vision and Memory", in Marcel and Bisiach (1992).

Whiten, A., ed. (1991). *Natural Theories of Mind*. Oxford: Blackwell.

Widerker, David (1995). "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities", *Philosophical Review* 104:247-61.

Widerker, David (2002). "Responsibility and Frankfurt-type Examples" in Kane (2002a).

Wimmer, H., and J. Perner (1983). "Beliefs about Beliefs", *Cognition* 13:103-28.

Winson, Jonathan (1997). "The Meaning of Dreams", *Scientific American* (November 1990); reprinted in *Scientific American Mysteries of the Mind*, Special Issue 7(1):58-64 (1997).

Wolf, Susan (1988). "Sanity and the Metaphysics of Responsibility" in Schoeman (1988).

Wolf, Susan (1990). *Freedom Within Reason*. New York: Oxford UP.

Wolf, Susan (1995). "Meaningful Lives in a Meaningless World". Machette Lecture. Brooklyn College of the City University of New York (Spring 1995).

Zaehner, R.C. (1966). *Hindu Scriptures*. New York: Dutton (Everyman's Library).

Zagzebski, Linda Trinkaus (2002). "Recent Work on Divine Foreknowledge and Free Will" in Kane (2002a).